Tampere University

Mohammad Ali Pourabed

DESIGN AND IMPLEMENTATION OF IDCT/IDST-SPECIFIC ACCELERATORS FOR HEVC STANDARD ON HETEROGENEOUS ACCELERATOR-RICH PLATFORM

Faculty of Information Technology and Communication Sciences
Master of Science Thesis
January 7, 2019

ABSTRACT

MOHAMMAD ALI POURABED: Design and Implementation of IDCT/IDST-Specific Accelerators for HEVC Standard on Heterogeneous Accelerator-Rich Platform
Tampere University
Master of Science thesis, 59 pages
January 7, 2019
Master's Degree Programme in Electrical Engineering
Major: Wireless Communications
Examiner: Prof. Jari Nurmi
Dr. Sajjad Nouri
Keywords: IDCT, IDST, HEVC, HARP, CGRA, Multicore, NoC, RISC, FPGA

Having High Efficiency Video Coding (HEVC) is important for image processing, reducing bandwidth, and increasing video quality. There are different methods that can be used to implement HEVC. This thesis focuses on design and implementation of application-specific accelerators for IDCT/IDST algorithms dedicated for HEVC standard. Those algorithms are parallel-in-nature tasks which makes them suitable to be executed by heterogeneous multicore platforms. This is done using accelerators which are required for power efficient processing. In this study, Coarse-Grained Reconfigurable Arrays (CGRAs) are used for making a template for an accelerator. CGRA has one of the major roles in a Heterogeneous Accelerator-Rich Platforms (HARP) as it is capable of accelerating non-parallel loops with lower loop counts. This thesis includes various algorithms for the use of IDCT and IDST with different designs and templates, reaching a unique final architecture. The final output intended is to reach 4 points IDST together with a 4/8 points IDCT. Another feature added to the hypothesis is the use of different dimensions for the CGRA template in order to have a different type of accelerator. The many CGRAs are combined together in successive arrangement with Reduced Instructions Set Computers (RISC) over the Network-on-Chip (NoC). The aim is to study the performance of the accelerator used for the IDCT and the IDST. This can be evaluated as the data movement through NoC network along with comparison of performance of accelerator with clock cycles in order to calculate the efficiency of the system. The results show that a four point IDST and IDCT can be computed in 56 clock cycles. In addition, the 8 point IDCT can be implemented in 64 cycles. One important factor to consider during the study is the power and energy consumption which is important in this century. The dynamic power dissipation usage for the routing of data has reached a value of 4.03 mW. Whereas, the energy consumption was 1.76 μJ for the 4 points

system (IDCT and IDST) and 3.06 μJ for the 8 points (IDCT). Processing Elements (PEs) are used for implementing the transform algorithm and units were operated at 200 MHz. Finally, these results show that 1080P image at 30 frames per second can be attained by using FPGA.

PREFACE

The basis for this research originates from my passion for understanding the High-Efficiency Video Coding technology. With the advancement in technology, it is necessary to have such technology which can support higher resolution and can allow people to experience the real impact of video coding. It is my passion not only to find out but also to contribute positively towards HEVC. This thesis is based on the research work carried out regarding the implementation of 4/8 point IDCT and 8-point IDST on HARP platform at the Tampere University, Tampere, Finland.

First of all, I would like to thanks my inspiration Prof. Jari Nurmi, who not only believed in my abilities but also allowed me to become a part of his research team. My Words can neither qualify nor quantify how helpful his guidance and his advice has been. I would also like to thank Dr. Sajjad Nouri who always remained a true guide during my research work and guided me with his valuable suggestions and feedback which allowed me to work on such a major project. Besides, Sajjad has become a brother to me as well given by nature. While I was in agony, he helped me out and built up my hopes. I truly believe that God sent him to me to wipe the tears from my eyes. I always love him, respect him and appreciate him.

I consider myself extremely fortunate to have been blessed with such great friendships during my stay in TUNI who have been extraordinary and benevolent. Special thanks to Naveen, Luis, Elena, Sun Bo, Dawood, Mehdi, Ritayan during those tough times in my life. And, I am always grateful to my dear friend, my flatmate Amir, who was always available for me and strengthened my courage. Besides, he always shelters me from the storm.

I am very thankful to my dear friend Dr. Ahmad Mardoukhi, who was giving me his valuable time during the long journey. Moreover, he is a truly supportive friend whom I really admired.

I would like to express my love to the dearest person in my life Katriina, who truly believed in me and made me risk everything for a future worth having. Whenever I disappointed, it was her love and memories which let me struggle more and more. During my entire life, her love was the best thing that has ever happened to me. She taught me how to love again. Her stunning eyes were like a home for me as a silent prayer. I truly believe that she is watching over me from the skies and completing this research work also made her happy in heaven. May God rest her soul in peace.

I devotedly thank my beloved sister who is a golden girl and a goddess. She always teaches me how to be strong, motivated and stay focused on my aims. I would like to thank Amin who always supports me in rain or shine. I would like to express my love to my lovely nephew, Satrap.

Finally, I extend my greatest gratitude to my parents for their true love and valuable support throughout my life and letting me to explore such a beautiful world. I owe my all achievements to them because without their support, I would not have been able to attain what I attained now. I am serendipitous to have such exceptional parents.

Tampere, 25.04.2019
Mohammad Ali Pourabed

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND SYMBOLS

ASIC	Application-Specic Integrated Circuit
ALM	Adaptive Logic Module
ALU	Arithmetic and Logic Unit
CC	Clock Cycle
CGRA	Coarse-Grained Recongurable Array
DSP	Digital Signal Processing
eFPGA	Embedded Field Programmable Gate Array
FCS	Feedback Control System
FFT	Fast Fourier Transform
FF	Flip Flop
FPGA	Field Programmable Gate Array
FU	Functional Unit
GOPS	Giga Operations Per Second
GPP	General Purpose Processor
HEVC	High Efciency Video Coding
HLS	High-Level Synthesis
I/O	Input/Output
LUT	Lookup-Tables
MFC	Multi-Function logic Cells
MOPS	Millions of Operations Per Second
MPEG	Moving Picture Experts Group
MPSoC	Multi-Processor System-on-Chip
NoC	Network-on-Chip
RAW	Recongurable Architecture Workstation
RISC	Reduced Instruction-Set Computing
RPISO	Reordered Parallel-in Serial-out
SDR	Software Defined Radio
VHDL	Very high-speed integrated circuit Hardware Description Language
VLIW	Very Long Instruction Word
COFFEE	Core For Free
DCT	Discrete Cosine Transform
DRF	Data Register File
DST	Discrete Sine Transform
HARP	Heterogeneous Accelerator-Rich Platform
IDCT	Inverse Discrete Cosine Transform
IDST	Inverse Discrete Sine Transform

MFC	Multi-Function logic Cells
PAE	Processing Array Element
PE	Processing Element
PU	Processing Units
RD	Rate-Distortion

1. INTRODUCTION

Since the inception of video technology, considerable amount of research papers have been published regarding the solving of problem of poor quality. Initially, the picture quality of the video was not good and the first video technology did not have the capability to provide voice along with the video. It was limited to video only due to its poor storage. However, with the advancement in technology, detailed research resulted in advancement in video technology and became the basis of the high-end technologies which include Inverse Discrete Sine Transform (IDST) and Inverse Discrete Cosine Transform (IDCT) [71]. The HEVC (High Efficiency Video Coding) which is commonly termed with the standard H.265 is considered as one of the innovative International Standards of the video due to its tremendous advantages [71]. HEVC allows the minimization of bit rate to 40 percent which not only increases the storage capacity but also enhances the transmission requirements of advanced video applications [71]. It is due to HEVC which makes it possible to access 4K videos which take a lot of space and make it difficult to stream if HEVC is not available [68]. Nowadays, , a number of 4K videos are available on different platforms with higher pixel. The advanced coding structure allows to have a good storage capacity and same is the case with the HEVC. It has an advanced coding structure which uses coding tree units. The coding tree units have the capability to support high-resolution pixel of 64×64 which is superior when compared to 16×16 pixels of H.264 [71]. However, HEVC also encounters various problems. According to a recent research study carried out by the Moscow State University in Russia, the performance of the HEVC was outpaced by the performance of slow mode of VP9. If the working principle of HEVC is considered, it can be analyzed that IDST and IDCT are utilized in order to simplify the matching between decoders and coders [71]. Transforms such as IDCT and IDST are specifically used for processing of the digital signals, MPEG, JPEG, and H.26x formats. Although IDST and IDCT have the capability of supporting a large range of block sizes, the problem occurs in case of compression of larger blocks such as up to 64×64 in case of HEVC; it results in complexity of computation and algorithm which ultimately decrease the performance [71]. As HEVC standard has been devised to support higher pixels and the slower performance in case of compression of higher blocks results in various

changes which can be carried out in order to address the issue. Recent research studies have proposed the idea of the modification of architecture of IDCT to be multiplication free as it would minimize the low hardware utilization along with the reduced access to peak bandwidth at the time of processing larger blocks [71]. Additionally, other research studies have researched other ways of decreasing the cost related to hardware along with power consumption by confirming that architecture of the IDCT decrypts the Ultra High Definition as well as Quad Full HD. With help of the multiplication free structure, the execution time of the decoding UHD videos can be increased to 30 fps [71]. Ultimately, it also results in lower power consumption and lower hardware costs to 25%.

At the time of modification of architecture, low hardware utilization is one of the major issues which increase the cost and one of the major issues faced by the developers. Previously, there were few transistors available on the integrated circuits and they were not sufficient In decompression of larger blocks. Now, ICs have transistors measured in billions and adding new transistors not only increase the cost but also impact the power usage. Power usage is related to heat dissipation and for each watt, a joule of heat is dissipated [67]. Though, the amount of transistors can be increased but the amount of power has not decreased which makes it difficult to work. So, in spite of the fact that a large number of transistors can be used on the chip, a major portion of the circuit cannot be used, which makes it difficult to cope with lower hardware utilization [67]. The complete circuit of the integrated circuit cannot be used and the remaining silicon that should be left unpowered is termed as the dark silicon. There were various changes which have been carried out in order to address the dark silicon issue. For carrying out the coding of larger blocks, it is necessary to address the challenge of dark silicon.

The architect chosen for this work is the Heterogeneous Accelerator Rich Platform (HARP). The design of the HARP comprises nine nodes which are organized in three columns and rows [71]. The central node of the HARP contains the COFFEE RISC core that has the functionality of the monitoring node but it also performs its role in general purpose processing. Other nodes of the HARP are either RISC processor or CGRA [71]. HARP is designed by changing the template-based CGRSs with the help of logarithm [71]. After modifying the HARP, the nodes are set up around RISC core and functions as the regulating device. It also aids the distribution of the data and the configuration streams so that handling of the slave nodes is integrated to the Network on Chip [71]. CGRAs are the capable accelerators as they are considered power-efficient accelerators with an array of Processing Elements (PE) which is connected with the help of the 2-D network [66]. Every PE has the ALU type Functional Unit (FU) and the Register File (RF). Functional Units have the

ability to execute memory, logical, and arithmetic operations. With each instruction cycle, every PE gets the commands from the instruction memory and specifies the operation [66]. The PE has the ability to write and read the data from memory, and data buses are shared by PEs in the same columns or PEs in the same row. CGRA has the ability to attain higher power efficiency due to simple hardware and efficient software techniques. The processing elements can be classified in two way i.e. homogeneous or heterogeneous [76]. Every heterogeneous processing element has a different instruction set while on the other end homogeneous has the capability to perform same set of instructions. While if the network in CGRAs are considered, it can be analyzed that there are two major types of network in the CGRAs which are multistage network and crossbar network [76]. Logical elements are less in case of multistage network while in case of crossbar network, the mapping is complex and difficult to implement it [76]. The crossbar network allows mapping from the inputs to any output which helps to ease the process of mapping and utilize major number of logical elements.

The aim of the research is to integrate 4/8 point IDCT and 4 point IDST on HARP template using CGRA. As explained earlier, the HARP is a multicore design at TUT and has shown results in terms of its utilization as the common purpose energy wave transceiver medium of different applications of IOT and to solve issues related to Dark Silicon. This is one of the major reasons for its usage to sort out the problems concerning Dark Silicon. In order to carry out the HARP testing, IDCT test has been used because it can be parallelized and as it is a computation-intensive task, it would be best for HARP testing. The aim of the research is to implement 4 and 8 point IDCT and 4 point IDST which are dedicated on HEVC standard with the implementation of Coarse-Grained Reconfigurable Arrays as the template based accelerators on HARP.

1.1 Thesis Outline

This thesis has been divided into 6 chapters. The Second chapter presents the detailed literature review regarding reconfigurable devices. Additionally, various state-of-the-art multicore platforms have been reviewed from the literature. The chapter considers the literature review regarding implementation of HARP. Chapter 3 explains the architecture of CGRA and HARP. Chapter 4 presents the design and implementation of 4/8 point IDCT and 4 point IDST using template based CGRA. Estimations, calculations, evaluations, and comparison of results have been discussed in chapter 5. Chapter 6 provides the conclusion regarding the implementation of 4/8 point IDCT and 4 point IDST on template based CGRA. At the end of the chapter, future work has been discussed which can be carried out to expand the research work.

2. LITERATURE REVIEW

Different type of application-specific accelerators have been developed in form of processor/co-processor model and embedded on the Multi-Processor System on Chip termed as (MPSoC) for carrying out the computationally intensive tasks [1]. There are different classes of accelerators and one of the widely used class is CGRAs, which have the functionality of acting as the co-processor to the processor in order to form the heterogeneous multicore platform in which both processors can be utilized simultaneously or can work independently [1]. Previously, single core has been used for carrying out various tasks before the creation of multi-core platforms in processor/co-processor models. Additionally, few accelerators have also been designed for carrying out computationally rigorous tasks. After the creation of the multicore platforms, VLIW machines have also been designed and developed for supporting the large-scale parallel applications [2]. With the passage of time, the architecture of VLIW is combined with the digital signal processors for tackling the DSP applications carrying out high-end mobile communication [3]. In the case of multicore platforms, the accelerators have the ability to operate as a co-processor in case of tight and loose coupling. In the case of tight coupling, higher bandwidth is used, which allows faster data transfer and synchronization as compared to loose coupling ([4], [5]). However, in loose coupling the accelerators are attached to the processor with lower bandwidth [6]. With the employment of co-processor bus or integration of the accelerator in the data-path, the accelerators can be coupled tightly to the processor.

2.1 Reconfigurable Devices

In previous years, the reconfigurable devices have been recognized as the popular hardware architecture due to the flexibility to carry out changes along with the reduction of cost and time in the development of systems. These devices have the ability to change their functions simultaneously on basis of their data flow indicated by the developer at the time of designing for carrying out various tasks [1]. Generically, there are three major types of reconfigurable devices which are recognized on the basis of their granularity, fine-grained having granularity of 4 bits or less, middle

grained devices having the granularity of less than or equal to 8 bits, and coarse-grained with the granularity of higher than 8 bits[7]. From these reconfigurable devices, the fine-grained devices are considered as having the most optimal resource utilization due to the presence of fine level granularity. While In middle grained devices, they are termed as the compromise between fine grained and coarse grained as they have the ability to process higher bandwidth [1]. While Coarse-grained devices are considered as having the simplest compilers and the higher level of granularity that supports a number of applications. The next section of the chapter will present examples of fine, middle, and coarse-grained devices from the literature. Additionally, the generic introduction regarding DCT will also been presented. The last section of the chapter will discuss the implementation of IDCT on various platforms which have already been discussed in the literature.

2.1.1 Fine-Grained Devices

Fine-grained devices have the granularity level of the processing elements from 1 to 4 bits and have more processing elements when compared to the coarse-grained devices. In today era, majority of applications is operating at 8, 16, or 32 bits which makes fine-grained devices of less interest for developers [1]. When compared with the coarse-grained devices, fine-grained devices employ more number of processing elements for executing the same operation and cost more resource utilization and poor mapping. There are various devices which are based on fine-grained operations and one of the most promising is the FPGA and particularly embedded FPGA (eFPGA) [1]. The architecture of FPGA consist of Logic Elements (LEs), Look-Up Table, 2 to 1 multiplexers that contain logic gates and Flip Flops (FFs). Xilinx [8] and Altera [9] are the most recognized fine-grained devise in the market. For example, the research study [10] involved the integration of three eFPGAs with the NoC-based system. Another fine-grained devices is GARP architecture which have the ability to act as a reconfigurable coprocessor, coupled tightly with the GPP and offering the lower granularity by 2-bit LUTs [11]. Fine-grained Device GARP have the PE arrays and each row consist of a single control block along with 23 logic blocks. At the time of designing, the size of the PE arrays can be increased or decreased on the basis of the requirements. For keeping the fixed operating frequency, the connectivity is limited on its fabric [12]. Other example of fine-grained device is FlexEos which works as eFPGA [13]. It is comprised of 4K Multi-Function logic Cells (MFC) on the basis of the SRAM 1-bit Lookup-Tables. While FlexEos (Reprogrammable SRAM based scalable FPGA fabric) developed on the higher concentration multi-function logic cells can be programmed by using the standard description languages for example Verilog and VHDL [1]. Another example

of fine-grained device is MOLEN which also has the ability to act as a coprocessor to GPP ([14], [15]). MOLEN can be mapped on the Xilinix FPGA chip while FPGA is acting as the accelerator. Despite MOLEN is separate from GPP physically, special instructions can be executed on it due to the ISA of the GPP.

2.1.2 Middle-Grained Devices

The concept of the middle-grained devices was introduced for supporting the world length up to 8 bits. Thus, only that algorithm can be mapped which have processing world length up to 8 bits. Mapping the algorithm on the middle-grained reconfigurable units is difficult as compared to the fine-grained devices due to the fact that it has the increased processing word length [1]. The middle-grained devices is considered as the good compromise among power, performance, and area along with the supporting of various word length applications up to 8 bit. There are different devices which are based on middle-grained reconfigurable devices and the example of it is PiCoGA-III which consist of Reconfigurable Datapath Unit ([16], [17]). The composition of each RDU has ALU of 4 bits, LUT with 4 bit, and 4-bits integer along with Galois field multiplier. Another example is DART which has the ability to support 8 and 16-bit processing word length ([18], [19]).

2.1.3 Coarse-Grained Devices

Coarse-Grained Devices are considered as one of the most promising platforms that have the ability to support 8, 16, and 32-bit arithmetic on a single processing element. In CGRAs, the array of the predefined processing elements delivers the higher level of granularity, higher computational power, higher data level parallelization, higher throughput processing, lower energy consumption and larger bandwidth [1]. They are programmable and reconfigurable with a higher level language and can produce an increase in performance while operating at a lower frequency [1]. CGRAs are suitable for carrying out the huge intensive signal processing due to the level of granularity and internal structure. They are considered as one of the best platforms for a number of applications such as for video and images processing ([24, [25]), Wideband Code Division Multiple Access (WCDMA) cell search [20], FFT [21], Correlation [22], and Finite Impulse Response (FIR) filtering [23]. Despite the fact that it provides all the advantages, it has higher transient power dissipation and yields a larger area of a few million gates. Additionally, the majority of CGRAs have a fixed set of processing elements which are not optimal for performance and cost [1]. There are numerous different CGRA architectures which are in use and have been described in the following sections of the chapter.

BUTTER

BUTTER has the functionality of acting as coprocessor for the COFFEE RISC core and was developed for carrying out the computationally-intensive tasks [26]. It has 48 array of processing elements. Yet, the size of processing element arrays can be increased or decreased at design time. The PEs are interconnected to each other in node to node fashion for carrying out the information exchange. The processing element has a functional unit for logic and arithmetic operations with fixed granularity at 32 bits and has the ability to support single precision floating point and integer [1]. The processing data and the configuration data can be transferred from main memory to the CGRA with the help of using a DMA device. The DMA device provides integration between data memory and CGRA. In H.264, BUTTER was instantiated for carrying out the plotting of 2D low pass-image filter as well as de-blocking filter [1]. It provides the selection of connections at runtime and it is characterized by the run-time configurability. The complete BUTTER platform is synthesized on an FPGA device.

ADRES

ADRES (Architecture for Dynamically Reconfigurable Embedded Systems) acts as a CGRA architecture attached tightly with the Very Long Instruction Word processor ([27], [28], [29]). It has numerous advantages as compared to other CGRAs and it shows increased performance, lower communication costs, simpler programming model, and significant resource sharing. The CGRA and VLIW are combined on the single architecture which has the two virtual function views, the reconfigurable array view and VLIW view. The architecture of ADRES consists of 8×8 elements reconfigurable array [1]. Its elements are arranged in a special manner which specifically includes Functional Units, routing resources, and Register Files. It has the first row of reconfigurable arrays as Functional Units and the remaining rows consisting of RFs and FUs [1]. These rows belong to the second view. The Functional Units have 32 bits data bus and can be heterogeneous associating various operations. They are combined together with the one multi-port global Data Register File. The RCs (Reconfigurable Cells) interact with the help of the multi-port global DRF, assigned connections between FUs, and Local Register Files [1]. For storing the intermediate data, the RFs can be engaged in such a manner that the words of 16 bits are stored in the local RF and 64-bits worlds are stored in global RF. The routing resources are built with buses, networks, and wires. The functionality of RCs is to speed up the data flow In parallel computing. While In execution of non-kernel

codes, VLIW is used with the help of Instruction-Level Parallelism [1]. As ADRES acts as the coprocessor, reconfigurable arrays and VLIW have the option to share resources which results in never overlapping at execution time. For generating the instances based on ADRES, the XML-based architecture language can also be used. ADRES is manufactured on a 90 nm CMOS technology and showed execution of 40 MOPS/mW [1].

Morphosys

The architecture of MorphoSys is designed for operating on 16 or 8-bit data ([30], [31],[32]). It is built of an 8×8 array of reconfigurable processing units termed as Reconfigurable Cells having configuration memory, higher bandwidth memory interface, and 32-bit general-purpose processor coupled tightly. RISC core guides the operation of the RC array. The RC is divided into four divisions. The data transfer can be started between the RC array and external memory by the RISC core with the utilization of the two sets of Frame Buffers each having two memory banks [1]. Every RC has the ALU for carrying out the fixed-point operations, multiplier, input multiplexers, shift unit, and register file. The configuration of RC array can be carried using a 32-bit context word which can be further distributed to every RCs in same column or row [1]. Additionally, addition of special instructions have been added to TinyRISC's ISA for transferring the RC array related operations. The operations are control operations, data and configuration transfer between main memory and the array [1].

PACT-XPP

PACT-XPP is centered on the graded array of the coarse-grained architectures ([33], [34]) and serves as a self-reconfigurable processing engine. It is comprised of 3×3 adaptive computing components (Processing Array Elements) and packet-oriented communication system. It has the partial reconfiguration capability and allows PAEs to work independently which implies that a few PAEs can be again reconfigured for carrying out the new functionality while other PAEs can execute the computation of data simultaneously. Special events signals initiating in the array can trigger the reconfiguration [1]. The mapping can be carried out with the C subset program with the utilization of vectorizing C compiler XPP-VC [1]. It produces the maximum performance of 57.6 GOPS at 150 MHz frequency [1].

2.2 Multicore platforms

Multicore platforms can be of the heterogeneous or homogeneous type. In the case of the homogeneous multicore platform, numerous RISC processors are joined loosely with one another while In heterogeneous, RISC processors are coupled tightly with reconfigurable architectures [1]. The code is written mostly in C In homogeneous platforms and can be spread equally to every RISC cores. In contrary, in the case of heterogeneous platforms, extra effort is necessary for programming the coprocessors and processors with the utilization of customized tools. In the case of the proposed research, the multicore platform HARP is used [1]. There are also a few other multicore platforms which have been discussed in the next sections of the chapter. These multicore platforms also exhibit similar properties and features.

2.2.1 MORPHEUS

It is considered one of the heterogeneous multicore platform accelerator ([36], [37]). It has the complex structure and dynamic reconfigurable SoC primarily consisting of three major types of reconfigurable devices [1]. These are fine-grained embedded FPGA, middle grained, and coarse-grained array which helps to lessen the power consumption. Basically, it is designed for heterogeneous digital signal processing in order to carry out the dynamic reconfigurable computing which is centered on the 64-bit NoC [38]. In MORPHEOUS, ARM 926EJ-S RISC processor is the master node which is assigned to control the communication, synchronization, and reconfiguration mechanism. The complete system has a detailed infrastructure which includes memories and communications for enabling the regularity between heterogeneous accelerators [1]. In order to provide efficient utilization, the platform has special software which not only contains the designing tools but also operating systems. The fine-grained device is FlexEOS as mentioned above. In the case of middle grained devices, the device is DREAM which is reconfigurable DSP core [1]. It has the 32 bit RISC core along with PiCoGA-III reconfigurable data-path which acts as the matrix of reconfigurable logic cells. It provides the performance of 0.2 GOPS/mW in a 90nm CMOS technology [1].

The coarse-grained device is XPP-III which is combined into the data path of a VLIW processor. It is designed for highly corresponding processing performance for spilling applications. All the reconfigurable devices exchange data among each other with NoC except the system modules. While Heterogeneous Reconfigurable Engines, I/O peripherals, and memory units are system modules. The complete MORPHEUS chip provides the performance of 0.02 GOPS/mW while developed on the 90nm CMOS technology with the normal active power of 700 mW [1]. The

delivering capability of MORPHEUS is 120 GOPS with the utilization of 90-nm technology for attaining the video observation motion recognition application having the power consumption of 2.5 W [1].

2.2.2 P2012

It is the power and area efficient core computing platform comprised of four clusters interacting with each other with the utilization of higher performance fully-asynchronous NoC [39]. The composition of each cluster is of 16 general purpose processors having autonomous instruction streams and the knots are generically locally synchronous and globally asynchronous. The communication between software and hardware is carried out with the utilization of the local and global interconnection, which act as the point to point stream communication [1]. In the case of P2012, the special hardware is dedicated to performing the synchronized and advanced power management. While the extended version of P2012 is termed as He-P2012 and can also be classified as the MPSoC platform. This platform shows the performance of the 40 MOPS/mW with the utilization of 28 nm CMOS technology [1].

2.2.3 NineSilica

NineSilica was developed at TUT by a research group for general purpose homogeneous MPSoC and having capability of programming in C language [35]. If the composition of NineSilica is considered, it consists of nine homogeneous cores, which are connected over the NoC in 3×3 mesh topology. In it, each node has the 32-bit COFFEE RISC processor [1]. While the center node has the working of supervision node for examining the other nodes. Every node has its own data memory and instructions. While the data can be switched in every nodes over the NoC with the help of the packet switching technique [1]. For testing the functionality of multicore platform NineSilica, numerous SDR applications have been implemented such as FFT and correlations. The results of the study revealed that 64-point FFT can be executed with the help of NineSilica in 10.3 microseconds on the FPGA device [1].

2.2.4 RAW

Multicore platform reconfigurable Architecture Workstation (RAW) consists of 16 32-bit modified MIPS2000 processors, which are organized in the array of order

4×4 mesh over the NoC [40]. It allows the static scheduling, which is similar in performance to the reconfigurable arrays and active scheduling, which is the mechanism similar to multi-core systems for carrying out the network transactions [1]. In the case of RAW microprocessors, the issue of wire-delay is managed by the programmable NoC and showing the wiring channel operator to software.

2.2.5 Fulmine

The development of the Fulmine has been carried out as the extensively specialized multicore platform for applications based on IoT specifically the smart secure near sensor data analytics [41]. It has the 65 nm SoC, which is created on the firmly coupled multicore-cluster strengthened with the dedicated blocks for carrying out the computationally severe jobs. In case of Fulmine, 32-bit OpenRISC cores are the four enhanced engines, which have the ability to exchange data with the accelerator in an efficient manner due to the employment of memory sharing mechanism [1]. It delivers the performance of up to 25 MIPS/mW with the power consumption of 20 mW on 0.8V [1].

2.3 Related Work

HEVC is considered one of the best standards for video compression. Many research papers have already been published which discussed hardware implementation of DCT/DST for HEVC. Majority of the research papers have discussed the provided output by HEVC which showed a 50 percent reduction in bitrate on the specific video quality [1]. As similar to H.264/AVC, the coding scheme of the HEVC is also hybrid block-based and includes intra and inter-picture forecast tools. In order to carry out the transform for each block of $N \times N$, the 2-D transform coding operation is implemented in such a manner that N-point 1D transform is carried out to each row and block separately. HEVC standard supports various transform sizes such as 4×4, 8×8, 16×16, and 32×32 Discrete Cosine Transform along with the 4×4 Discrete Sine Transform [1]. As it provides higher transform sizes, an additional bitrate reduction of 5% to 7% is achieved as compared to the conventional transform which is carried out in H.264/AVC. Although such transforms In HEVC showed performance in the Rate-Distortion (RD) but the complexity increased enormously [1]. In the design section of the paper, it can be analyzed from the design and implementation of the 4 and 8 point IDCT and 4 point IDST which is dedicated for HEVC on HARP template [1]. There are numerous research studies in the literature in which different transform has been carried out and presented in the following part of the

chapter. In a research work in [42], the high-speed two-dimensional IDCT processor for the video coding has been designed in which the processor used the row-column approach for calculating the 2-D IDCT in a manner that the complete architecture is separated into 1-D IDCT calculation with the help of a transpose buffer [42]. In this case, the 1-D IDCT scheming is carried out with help of the Loeffler algorithm and the process which involved multiplications is carried out with additions and shifts. The pipelining is presented for designing the circuit so that data can be disposed of in the equivalent manner. The concept of Loeffler algorithm is introduced for gaining a higher operating frequency [42]. This case study also introduced the row preprocess module which was developed to dispose such rows which have zero input. The introduction of the row preprocesses module helped to increase the decrypting speed of the 2-D IDCT processor. 5015 logic elements of Altera EP2C20F484C7 FPGA are used by the processor and gained the operating frequency of 117.37MHz [42]. Another research study [43] proposed the 4/8/16/32 Point Integer IDCT architecture for different video coding principles [43]. The proposed architecture had the capability to support various video standards such as MPEG-2/4, AVS, and HEVC [43]. In this research study, multipliers MCM were used for carrying out the 4/8 point IDCT while normal multipliers were used for 16/32 point IDCT. For reducing the hardware, the transpose memory used SRAM. Real time-video decoding of $4K \times 2K$ with 18944 SRAM and 93K gate count is carried out [43]. The 5 stages pipeline architecture is enabled in this research study too for attaining the higher working frequency but it also resulted in an increase in silicon area. Authors in [44] presented the high-performance 2-D IDCT for decoding of video which is centered on the FPGA which also used the same methodology as it was carried out in [1]. This design is comprehended in Xilinx Vertex5 Field Programmable Gate Array (FPGA) (44). The 2-D IDCT compressor has the higher accuracy, lower complication, and augmented speed. The advantage of using Loeffler's fast algorithm is that it helps to reduce power consumption. This research study is an extended version of [1] which improved the Loeffler's algorithm and attained a higher level of working frequency and higher accuracy IP. Additionally, the parity of Loeffler's algorithm is used to reduce the difficulty of the process. This paper also proposed the competent pipelining FPGA employment of the 2-D IDCT decoder which helped to attain the frequency of 278 MHz [44]. The implementation of the row-column approach helped to simplify the multiplications. The pre-processing module included in the research study included two major parts i.e. sequential conversion into parallel and zero-value judgment [44]. In another research work in [45], the reconfigurable IDCT architecture on FPGA for different video standards has been designed. It is used in the multi-standard decoder of VC-1, MPEG-4, and MPEG-2. The architecture included two-circuit sharing strategies, factor share along with adder share in or-

der to save the circuit resource. The research study used the Recursion property of DCT transform in order to solve the issue of numerous multiplications and additions [45]. The multiplier less transform is preferred as each element is uttered as the total of the different binary factors. For increasing the circuit utilization, factor-sharing strategy is used which helped to optimize the circuit and with the help of FS, numerous adders and multipliers were saved [45]. All type of 8-point IDCTs is divided into the 4-point IDCTs T4 along with 4-point IDCTs V4, permutation matrix P8,r, and the butterfly matrix P8,1 [45]. The used architecture in the research study was of low-cost and efficient circuit sharing is carried out on the basis of AS and FS strategies [45]. Another research study [46] considered the hardware-scheme for the 32×32 IDCT of the HEVC video coding standard [46]. This research study also utilized the inverse discrete cosine transform with the help of video encoder and decoder. The principle used in the paper is of separability. It was scheduled to reach the real-time dispensation of a minimum of 30 frames per second for higher resolution of video and exploiting the higher level of parallelism i.e. 32 samples per clock [46]. It was designed on the combinational way and with the help of the multiplier less approach. The synthesis was directed to the Altera Stratix IV FPGA. According to the results of the study, the architecture was able to process more than 30 QFHD frames along with the latency of 33 clock cycles [46]. The design was divided into five major parts which include two-registers set, one transposition matrix, and two 1-D IDCT architecture. The 32 points IDCT design used the two occurrences of the 1-D IDCT in order to explore the separability process. In the first part of the intended 1-D DCT, the design handled the multiplications and the process of multiplications was further decomposed into shifts and adders as discussed above in another research study which also employed the same methodology [46]. While the subsequent part of 1-D IDCT design executes the butterfly operations in which calculations and additions were carried out. The results of the research study were synthesized on the EP4SE820F43I4 device [46]. The design of the 32 point IDCT helped to attain the lower latency, higher processing rates, and lower hardware utilization. The higher dispensation rate was attained with the help of parallelism exploration and lower latency is attained with the help of the composite design in the 1-D DCT transforms [46]. While lower hardware cost is attained through the multipliers approach and decomposing the process of multiplications in adds and shifts. Another research work [47] designed the 2-D adjustable block size IDCT design for HEVC standard with the help of block size scheduling scheme which supported the variable blocks of various sizes such as 4×4, 8×8, and 32×32 pixels [47]. In this research study, TSMC 65nm 1P9M technology was used to synthesize the results and the results of the study showed that the 2-D design attained the higher work frequency of 400 MHz with the cost of hardware up

to 112.5K Gates [47]. The recursive and normal butterfly calculation arrangement is unfolded which helped to tackle various block sizes for IDCT. This research study also employed the customary row-column method and the design employed the 1-D Column Transform Core, 1-D Row Transform Core, and the Transpose Memory [47]. The transform cores used in the paper have a similar structure but have different data width. While the architecture of 1-D Transform Core adopted the 1-D linear systolic array architecture which included Array Units. These Array Units included the Delay Unit and two IDCT elements [47]. Another research study [48] considered the algorithm of 8×8 IDCT for HEVC. The proposed algorithm in the research study showed 66 percent fewer multiplications and 46 percent fewer additions when compared to the traditional method and it also saved 60 percent area for implementation of hardware [48]. The algorithm is also illustrated with the help of the signal flow graph which is easier for implementation on software or hardware. For understanding the results of the study in a better manner, it was synthesized by Synopsys Design Compiler with the help of SMIC 130nm CMOS library [48]. This algorithm considered the coherence property of the integer cosine transform which allowed to split matrix into odd and even parts. These odd and even parts In 8×8 IDCT are further decomposed into sparse matrices [48]. Another research study [49] considered the power effective and high troughtput multi-size IDCT considering the UHD HEVC decoders [49]. This research study presented the hardware architecture which aimed to gain the real-time handling of 30 frames per second and exploiting advanced level of parallelism [49]. The architecture was developed in the combinational manner which included multipliers approach and employed the optimization algorithm with the help of actions reuse and sub-expressions sharing [49]. The technology used in the paper is Altera Stratix V FPGA and ASIC 90nm standard-cells technology [49]. This research study also employed the same principle implemented in other research studies i.e. division of 2-D IDCT into two matching 1-D IDCT units which helped to carry out the further calculations [49]. The first module in this research study is used to compute the multi-size 1-D IDCT as input and the input size can be according to the transform size applied [49]. After that, a transposition matrix is used for providing the properly planned inputs to the second 1-D IDCT module [49]. While the transposition matrix was executed with the help of a blank of registers monitored by multiplexers. The designed architecture was synthesized on the 5SGXMABN3F45I4 device and results of the study showed that it attained the results which were aimed [49]. In a research study [50], an area and throughput efficient 2-D IDCT/IDST VLSI design were presented for HEVC standard which adopted the data flow development and common constant multiplication structure [50]. The design helped to support various block sizes. With the help of 65nm technology, the synthesis results revealed that highest working frequency is 500MHz

and the hardware cost is 145.4K gate count [50]. The results of the study showed that the designed architecture helped to cope with the actual HEVC of $4K \times 2K$ at 30 frames per second video categorization at 412 MHz on average. In this research study, the projected 2-D IDCT design supports various block size IDCT operations and in every cycle, the remaining data is forwarded to the Specific Multiplication Array and Template Operation Unit for carrying out different operations in a parallel manner [50]. With the help of Product Switch Network Unit, the data from Multiplication Array is forwarded to the Accumulator Array Unit and the proposed architecture attained higher than 50 percent hardware cost decrement and 66 percent throughput efficiency enhancement [50]. While if the IDCT is considered, it can be analyzed that it is one of the best tools for processing of digital signals and it has a number of applications in the area of multimedia as discussed above. According to research [51] carried out on DCT and IDCT, the pipeline implementation is carried out on the basis of the perfect shuffle topology algorithm. First of all, the accuracy of the structure is analyzed with MATLAB for knowing about the requirements regarding internal word length for the implementation. After that, the structure is modeled as the data path structure with the help of Synopsys Module Compiler [51]. According to the results of the study, the pipeline showed the operating frequency of 253MHz and used 40000 gates [51]. For accuracy analysis, the fixed point arithmetic is used for area efficiency. Additionally, for the accuracy analysis, C-language is used for modeling the parameterizable simulation model of the pipeline structure [51]. Another research study [52] presented the hardware architecture of the 4 point IDCT inverse transform unit for HEVC [52]. The research study proposed a simpler method for calculating the HEVC 4-point IDCT. In this methodology, the focus is given to the occurrence of the special cases in which results can be obtained without having full IDCT processing. With this approach, the number of calculations for 1-D IDCT reduced to 87.5 percent and gained an increased rate of 1.4 percent of BD-Rate [52]. The main purpose of the project is to attain the present processing of UHD 4K video with lower hardware utilization and increased presentation. The system was employed targeting the Cyclone V FPGA device. The results of the synthesis revealed that the system has the ability to practice the UHD 4K videos with the processing of 100 UHD 4K frames per second [52]. Additionally, the reduction of hardware source is also carried out up to 72.3 percent [52]. The research study involved designing of architecture for implementing the Fast 2- D IDCT which comprised of 4 major shares. These parts are two register sets for input and output, one divider unit, and a single 1-D IDCT 4 point architecture [52]. While the design of the 1-D IDCT 4 points composed of Multiplications, Butterfly Block, and Rounding Stage [52]. In another research study [53], the FPGA implementation of HEVC Invest DCT is carried out using high-level synthesis. The IDCT transform algorithm

is responsible for eleven percent of the calculations intricacy of the HEVC video encoder. This research study used the first FPGA implementation of the HEVC 2D IDCT algorithm with the utilization of HLS tools [53]. The provided hardware is implemented on the Xilinx FPGAs with the help of three major HSL tools and these are Xilinix Vivado HLS, LegUp, and MATLAB Simulink HDL Coder [53]. The development time of FPGA is reduced with the usage these tools and attained an increase in the performance which implies that HLS tools can also be further used for FPGA execution of HEVC [53]. Xilinix Vivado HLS helps to generate the Verilog RTL codes from source and System C codes. It also optimizes the speed, area, and power dissipation [53]. While LegUp is the open source HSL tools which can produce the Verilog RTL codes from C codes [53]. It delivers loop unrolling and pipelining. While MATLAB Simulink is commonly used modeling tools for numerous applications [53]. It helps to generate the Verilog RTL codes from the Simulink models and provides numerous optimization options such as clock gating, RAM mapping and pipelining [53]. According to another research study [54] in which re-configurable 2-D IDCT design for HEVC encoder and decoder has been presented [54]. The research study proposed the new configurable pipelined architecture in order to carry out the Inverse Discrete Cosine Transform and the circuit supported all type of transform block sizes with reconfigurability and reusability. The circuit is implemented on the TSM 65 nm and run at 500 MHz clock frequency in order to attain the throughput of 1990 Mpixel/second which is higher than any other architecture [54]. The discussed architecture to process the UHD video and have the ability to support up to 8K with 60 frames per second [54]. The architecture presented in the research study has two major components i.e. transpose memory and 1-D IDCT. The memory has the role of intermediary between two 1-D IDCT units and function as the buffer unit for saving the output retrieved from the first IDCT unit [54]. The main features which were covered in this architecture are configurability and reusability. The architecture is also pipelined for gaining higher throughput. The transpose circuit comprised of the register and the multiplexer [54]. While the multiplexer chose the controls of the signal and the data written on the register and then transfer it to rows and columns [54]. The architecture presented in the research study used Verilog HDL and mapped to the TSMC 65 nm cell library with the utilization of the Synopsys Design Compiler. The gate count is 197K gate [54]. Video coding standard HEVC involves the increased computational complexity. Another research study [55] which presented the lossless IDCT design for AVS2 described another methodology which involved skipping of the calculation of zero coefficients. The research study carried out after numerous statistical analysis and different patterns for transform blocks having different sizes were designed in order to detect the non-zero coefficients [55]. According to the research study,

if there is a confirmation by the transformation block of the calculated patterns, the streamlined IDCT role will be performed by the system [55]. The results of the study showed that the devised design could help to reduce the computation by almost 19 percent under various conditions. Additionally, the methodology did not produce any coding performance issue [55]. The research study involved the test for measuring the possibility of the non-zero coefficients in the block with different QPs [55]. The proposed method in the research study started from the inverse quantization process and ended at the IDCT function [55]. After IQ, the next step involved is the location of the non-zero coefficients and then analyzed to find out which mode should be implemented. In the end, the corresponding transform is carried out in the research study. The detection of the non-zero coefficients helped to implement the fast IDCT design and allowed to save time up to 19.3 percent without any loss in terms of performance [55]. Another case study [56] presented the high-level synthesis execution of the integer discrete cosine transform and discrete sine transform for HEVC [56]. This research study implemented the 2-D transform with the help of two 1-D transforms using the Even-odd decomposition techniques and common row-column approach [56]. The implemented architecture carried out the 4 points IDCT/IDST for the transform blocks and used the transpose memory for intermediate results [56]. The design is implemented on the Arria II FPGA and helped to support coding of 1080 pixels at 60 frames per second and the hardware cost was 216 DSP blocks and 10.0 kALUTs [56]. In this research study, the DST and DCT algorithm is acquired from open source Kvazaar HEVC encoder and the proposed architecture implemented the hardware-oriented even-odd division algorithm and its C code is combined to HDL with HLS [56]. The HLS helped to reduce the design and verification time and outperforms the other approaches in terms of cost and performance. Another research study [57] also carried out the high-level synthesis execution of 2-D IDCT/IDST on FPGA and used the same approach as mentioned in the above research study. This research study also implemented the 2-D transform with the help of two successive 1-D transform with the utilization of the Even-Odd decomposition technique and in this research study, the study made use of the HLS to implement the architecture from the C code of the algorithm [57]. It was also implemented on the same architecture i.e. Arria II FPGA and supported 60 frames per second. But, it has better resource management and five times faster than other solutions [57]. This research study also reduced the arithmetic operations with the help of Even-Odd differentiation algorithm commonly termed as Partial Butterfly algorithm [57]. This research study also utilized the transpose memory and 2-D IDCT transform [57]. The designed architecture has the ability to support Ultra HD video encoding at 35 frames per second and 68 fps. The architecture supported the video decoding of 2160p at the expense of 12.4 kALUTs and

344 DSP blocks [57]. Discrete Cosine Transform (DCT) is such a tool that have a number of presentations and various purposes [58]. In the case of video encoding and decoding, it is considered one of the most commonly used too. Along with that, there are other tools which can also be used. The research study [58] which was carried out regarding the employment of DCT and IDCT for image firmness and decompression on FPGA revealed that it helped to discrete the image into important parts. The designed DCT core considered taking higher area optimization and process audio frames and images which 512 cycles to process the eight-bit words [58]. This research study has considered the implementation of the design in VHDL with the utilization of the Behavioral model [58]. The total memory utilization in this research study is 75488 kilobytes. Area efficiency is one of the major objectives of numerous researches revolving around the architecture designing for HEVC decoder. Another research work [59] carried out regarding the area effective 4/8/16/32- point IDCT design for HEVC devised the area reduction by reducing the computational logic of the 1-D IDCTs with reordered parallel-in-serial-out (RPISO) scheme which shared input of the butterfly structured and reduced the area of the transpose buffer with the cyclic memory organization which attained 100 percent I/O utilization of SRAMs [59]. For implementing the unified 4/8/16/32 point IDCT, the suggested scheme showed thirty five percent reduction in terms of logic cost and a 62 percent reduction in terms of memory cost. The IDCT implementation of the architecture supported real-time decoding of the $4K \times 2K$ 60 frames per second video along with the hardware cost of 357,250 um^2 on the 2-D IDCT and 80,988 um^2 on transpose memory [59]. This research study considered the RPISO scheme and used SRAM as an alternative of the register for implementing the transpose memory. There was a usage of four SRAMs as the data parallelism for 1D IDCT architecture is 4 pixels [59]. In every cycle, there are 4 IT1 grades written in the memory buffer [59]. The 100 percent I/O utilization for SRAM is carried out with the help of the cyclic memory organization method with every cycle and every I/O port is used for reading and writing. The research study proposed the method in Verilog HDL and synthesized with the TSM 90nm cell library [59]. This research study supported real-time decoding of 4K2K with 60 frames per sequence video sequence [59]. Hardware reuse is also a method to sort out the issue of huge computational complexity. In a research work [60] carried out regarding the large IDCT for HEVC, the issue of huge computational complexity is resolved with the help of hardware reuse. The processing elements are optimized with the help of making changes in the regular butterfly structure and fully recursive structure [60]. The processing elements are implemented without multipliers and with the help of adders and shifters. The implementation of the architecture is carried out on 0.18um technology and showed 300 MHz frequency and 287Kgates areas which allowed to process 4K videos at

30 frames per second [60]. This research study used Chens fast DCT algorithm in which input is treated with 8-stage butterfly operations. While the requisite density involves the implementation of the larger area which is addressed with the reusing of processing elements. They are optimized with the help of shifters and adders [60]. For solving the issue of large multiplexer size In PE architecture, the position lines of input values were analyzed from top to bottom. The proposed architecture consists of 2 processing elements and one transpose buffer just like other research papers. This research study also used a similar methodology and attained the results of lower hardware utilization [60]. A similar research study [61] was carried out regarding the designing of the low-cost hybrid design of IDCT for H.264 and HEVC [61]. This research study involved the advancement of the generalized decompose and share algorithm with the utilization of the symmetric structure and factoring the matrix into submatrices. After that, matrix decomposition is carried out. The research proposed the generalize algorithm and hardware joint design with the help of utilization of the symmetric property of integer matrices and matrix division [61]. The design has been carried out in such a manner that it can cope with all change at any stage. According to the results of the study, the design has all four codecs and attained the maximum decoding capability [61]. The low energy HEVC IDCT hardware is proposed in another research study [62] which decoded 48 quad HD video and reduced the energy consumption almost by 23 percent. This research study proposed a novel energy reduction technique in order to avoid the IDCT for zero coefficients [62]. While technique checks DC coefficients and 3 lower frequency coefficients in the TU [62]. If there are DC coefficients different and have values other than zero along with all three coefficients having a lower value than the threshold value, the devised technique performed the IDCT for the DC coefficients in the TU. If the condition does not meet, it would perform the IDCT for every coefficient in the TU. This research study also used the butterfly structure and the selection of IDCT inputs is carried out on the basis of the TU size. For reducing the number of adders, the Hcu MCM algorithm is used in order to calculate the IDCT matrices [62]. The proposed technique in the research study and architecture is implemented with the help of Verilog HDL. The code is charted to the XC6VLX550T Xilinx Virtex 6 FPGA [62]. While the FPGA implementation used almost 34344 LUTs, 32 BRAMs, and 13811 slice registers [62]. Another research study [63] also presented the area competent 4/8/16/32 point IDCT design for the HEVC decoder [63]. In this research study, the hardware cost was reduced in terms of two major aspects i.e. first of all, logical cost of 1D IDCT is reduced with the help RPISO scheme [63]. With the help of this scheme, number of calculations for inputs of butterfly were reduced in every cycle. While the second aspect of hardware reduction is that the area of transpose memory is reduced with the help of the cyclic data

mapping scheme which helped to gain 100 percent I/O utilization for every SRAM [63]. For designing the pipelined 2D IDCT design, the pipelining program for every column and row transform is carried out. According to the results of the study, the area can be reduced up to 25 percent for the rational IDCT and memory area can also be lessen up to 62 percent [63]. This research study is an extended version of a research study carried out regarding the Area efficient 4/8/16/32 point inverse DCT architecture for UHDTV HEVC decoder. In both research studies, SRAM is used instead of transpose memory for saving the area. Additionally, both types of research included the symmetric property in the butterfly structure and used the RPISO scheme for reducing the inputs of the butterfly and to lessen the number of calculations [63]. This research study is different from the above-mentioned research in a slight manner as in it, two-port SRAM is utilized and pipelining program is used for row and column 1D IDCT for avoiding the issue of writing and reading [63]. According to the results of the study, the projected design has the ability to support actual video decoding of $4K \times 2K$ at 60 frames per second [63]. With the help of utilization of Chens algorithm, the N-point design is reclaimed in the 2N-point design.

100 percent hardware utilization is difficult in an architecture compatible with HEVC. According to a research study [64] which considered the designing of the fully-pipelined 2-D IDCT/IDST VLSI design compatible with the HEVC, the 100 percent hardware utilization is possible and can be carried out [64]. It was implemented on the SMIC 65 nm 1P9M technology, the results of synthesis showed that architecture attained the extreme frequency at 480MHz and the complete hardware expense for it is 115.8K Gates [64]. While if the experimental results are considered, it can be analyzed that this design is also able to deal with actual video of $4K \times 2K$ at 30 frames per second at 171 MHz in average [64]. This research study changed the granularity of IDCT computation by unrolling the butterfly computation assembly for removing the correlation. The unrolling of the butterfly operation helped to change the granularity. While traditional row-column decomposition approach is used in this research study which included the 1-D column transform core and 1-D row transform core along with the transpose buffer unit [64]. The results of the research study revealed the total power of 56.36mw. The hardware overhead for 1-D column transform core and 1-D row transform core is 60.5K NAND2 gates and 55.3K NAND2 gates [64]. Various research studies have proposed different methodologies for hardware complexity. According to a research study [65] regarding energy and area effective hardware execution of HEVC inverse transform, the pipeline scheme can be implemented to process any transform size with lower throughput of 2 pixels with zero-column capering in order to improve the throughput [65]. In this research

study, another approach was used which involved data-gating in 1-D IDCT engine in order to improve the energy efficiency for smaller transform sizes [65]. Instead of using the transpose memory, this study also involved the utilization of SRAM based transpose memory in order to have an area efficient design. The designed architecture supported 4K videos at 30 fps and the hardware utilization involved 98.1 kgate logics and 16.4 kbit SRAM [65]. These are different research studies available in the literature regarding the architecture for HEVC standard. One of the interesting factors which can be analyzed from these research studies is that in the majority of research studies, 2-D IDCT has been carried out with the help of butterfly operation and it has been decomposed to two IDCT operations. The output from the first IDCT is input into transpose memory and then output from transpose memory is input to the other 1D IDCT. While in the presented research work, 4 and 8 point IDCT and 4 points IDST has been carried out.

3. PLATFORM ARCHITECTURE

3.1 Coarse-Grained reconfigurable Arrays (CGRA)

Although the modern HEVC techniques have improved drastically, they have their limitations. For instance, efficiency is often driven by an application's maximization of the limited computational resources available; the limited storage space and transmission speeds required. The CGRA is a template based co-processor design with each template being equipped with a multiple of rows by columns (R * C) of processing elements (PEs), which can be scaled depending on the application to be processed. The R is application dependent while the C of the template based CGRA can be scaled between 4 up to 32. To increase the core's efficiency, two local memories are packed with a maximum capacity to accommodate 32 rows by 512 columns. Inside the CGRA, for data to be distributed between the PEs and the local memory, tow I/O buffers are usually integrated onto the chip. The building technology of the I/O buffers is based on C with C being equal to the total columns of local memory, $C \times 1$ multiplexers and C 32 bit-registers. Each of the PEs within the template-based CGRA has an accompanying two inputs and two outputs. Moreover, the PEs within the architecture do have the LUT, adder, multiplier, Shifter, immediate register, along with the floating-point logic. All the additional elements can be used by a designer for instantiation at the design time. Flexibility is a driving point in the implementation of the PEs in the design. The PEs interconnect in such a manner to offer a designer enough flexibility to develop the connection amongst the neighboring PEs in the node to node fashion to offer different routing options. The connection can be globalized in its connection, can be localized or the PEs can be interleaved together [38] The structure of a template-based CGRA has much reliance on its capability to scale. Scaling up or down of a template-based CGRA has much reliance on the algebraic expressions driving the application in, which it is intended. It is governed by its R * C architecture, accompanying local memories, a set of I/O buffers, PEs and interconnection of nodes. Equipped with R * C topological arrangement, to determine the application to be used for, the R in the R * C design is application dependent with the C offering scaling capabilities of between 4, 16 or a 32 scaling. In order to enhance its memory efficiency, there are

accompanying integrated local memories with a capacity to accommodate a 32 rows by 512 columns on full capacity. Furthermore, data sharing between PEs and local memories is facilitated by a set of I/O buffers integrated onto the chip. I/O pair of buffers are of the C type design; a C1 multiplexer and a C 32 bit registers with C being equal to local memory's total number of columns. The node interconnection offers flexibility benefits to the designer to interconnect multiple PEs. The node design is a 3 by 3 design based with the innermost node being integrated with a RISC core architecture. Specifically, the inner core is tasked with overlooking the other nodes in that, besides being used as a supervising node, it can also be used in the processing of general purpose applications. The other outer lying nodes however, their design architecture can be based on RISC architecture or template-based CGRA [75].

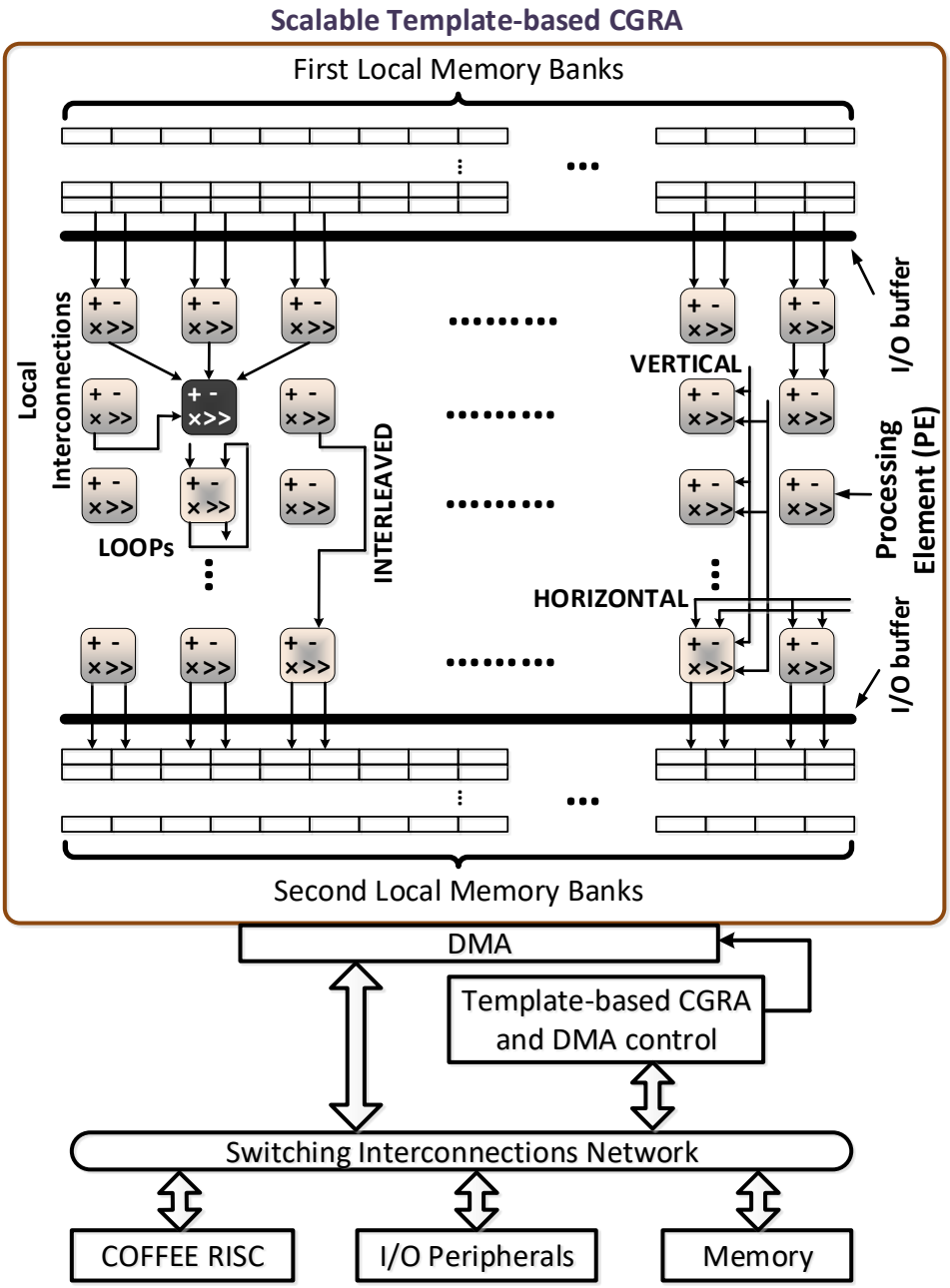

Figure 3.1 *The architecture of a scalable template-based CGRA used for integration over Network on Chip © 2018 IEEE [71]*

Data sharing between the local memory and the processing elements is facilitated by a pair of I/O buffers integrated onto the chip. Processing is performed by the PEs with the help of adders, multipliers, shifters, LUT, immediate registers and

floating point logic. During the design of the chip, the supporting components of PEs can be instantiated at the design time. The flow of data in the CGRA chip is governed by the algebraic expression of the input stream. Maximum use of the PEs has been carried out at the time of loading the algorithm in case of designing of chips. First, the configuration template has to be loaded from the outer source with help of DMA which is the direct memory access which loads the configuration file onto the CGRA. From the CGRA, data flow within the chip is governed by nodes and their interconnection using a set of I/O buffers to allow for data flow between the CGRA registers and the PE.

The PEs structure loaded via the DMA onto the template-based CGRA contains an address and an accompanying operations of the configuration stream. The addressing allows the PEs to determine the destination of the configuration file while the operations on the stream tells the PEs what to do with stream. Following which, the data which has to be processed is transferred from the RISC processor and then loaded onto CGRA processor for execution via local CGRA chip's local memories. The interconnection of the PEs are flexible. After the data, which has to process is added to the CGRA processor, the context has to be enabled within the processor to constitute the PEs functionality and the functionality between them. A designer has the liberty to alter the configuration of the PEs and its context stream to suit an application to be loaded onto the chip at runtime. As a result, the data in the CGRA can be managed by the PE and added in the second local memory or the new stream of data can be fetched at runtime of the process in the meantime. The processing structure of a CGRA processor allows the designer the flexibility to iterate the process of loading data onto the CGRA from DMA, through the local memory to the CGRA chip and eventually to the second local memory. The process flow has to be iterated until eventually, the process reaches its completion.

The general data flow of processes follow a stream of loading from a DMA, to the CGRA processor. From the CGRA processor, data flow is between process elements (PE) and local memory. A designer can instantiate a PE's adder, its multiplier, shifter, registers, and the floating point logic to handle an input configuration stream. The flexibility of PEs interconnection allows for a point to point connection to be made in a floating point fashion based on either a local, a global or an interleaved connection.

3.2 HARP (Heterogeneous Accelerator-Rich Platform)

Heterogeneous microprocessor design helps to integrate the multicore dissimilar processors on the processor chip for performing other functionalities. The HARP is a platform comprising 9 nodes arranged in sequences of 3 rows and 3 columns organized in a mesh design formation. The central node acts as a supervisor for the other nodes on the platform and being in unified with the COFFEE RISC core, it also can be used for general purpose processing. The other nodes on the mesh network can either be RISC processor. Heterogeneous platform allows for the addition of coprocessors to improve the performance of a processor. The template-based CGRA on the architecture can be scaled either to go up or to down with a major dependence on the algebraic expressions of the kind of application it is intended for. The design principal of HARP is centered on the CGRA processor core. Each of the nodes in a NoC architecture design has a master interface and two slave interfaces accompanying. The master interface of the node is combined with the RISC core or it is linked to the master side of the DMA device. As a master interface, it has the ability to write to network as well data transferring within the nodes of the architecture. While on the other end, the slave interfaces on the other hand, unlike the supervisor interface, are combined with the template-based CGRA or with the slave side of a Direct Memory Access (DMA) with their sole purpose being the reception of data from the NoC.

The implementation flow of data inside the HARP architecture, begins with the loading of a configuration stream through a direct memory access. Then, the execution words of the configuration stream together with its accompanying data is loaded from the supervising node of the processor onto the slave node in form of packets. Conventionally, the packet comprises two adjoining parts; the header and the data and configuration words. The header contains the routing information; the source address and the destination address. To enhance its efficiency with consecutive data transfers routing to the same slave node or transfer of data between different nodes, to avoid the smash of data, the master interface has to establish a synchronization mechanism. Master node synchronization is achieved by setting and resetting the read and the write flags of the allocated shared memory that does correspond to the destination node. To achieve the seamless synchronization, the master node writes a 1 onto the shared memory location.

Once the DMA transfer of data is complete, the DMA master sends an acknowledgement to the supervisor node, which then writes a 0 to the shared memory location on the NoC. In turn, writing a 0 indicates the complete transfer of data therefore resetting the shared memory location and releasing it to be used by other processes. Once data has been loaded from the DMA onto the template-based CGRA local

memory, processing can be handled by PE arrays and the output data stored on the second local memory. Should the CGRA based on template complete the processing of data, the processed information can be transported back to the data memory of the RISC processor core or can be drew directly by the local memory of another template-based CGRA for additional dispensation. As the data undergoes processing,. To achieve accuracy in their output, the RISC core sends control words to the template-based CGRA. The working performance of CGRA core processors can be parallel yet independent of each other or can be parallel yet dependent on each other such that data exchange can occur between the CGRAs in the event of a data dependent program flow [74].

The process of data loads into the NoC follows a systematic process. The configuration file has to be loaded using the master interface of a DMA onto the supervising node of a HARP architecture processor. From there, data processing decisions lies with the COFFEE RISC based supervisor nodes to tell the salve nodes, which node to perform the processing and what operations to perform on the data. To achieve accurate data processing, the supervisor node sends control signals to the slave nodes. Inside the CGRA co-processor, the local memories hold data to be processed by the process elements with the output being loaded onto the second memory chip. Access to, which however, can be from a slave interface of a DMA or other CGRA processors should the data need further manipulation. Synchronization is achieved by the control of the master interface setting and resetting data flow in a shared memory location [73].

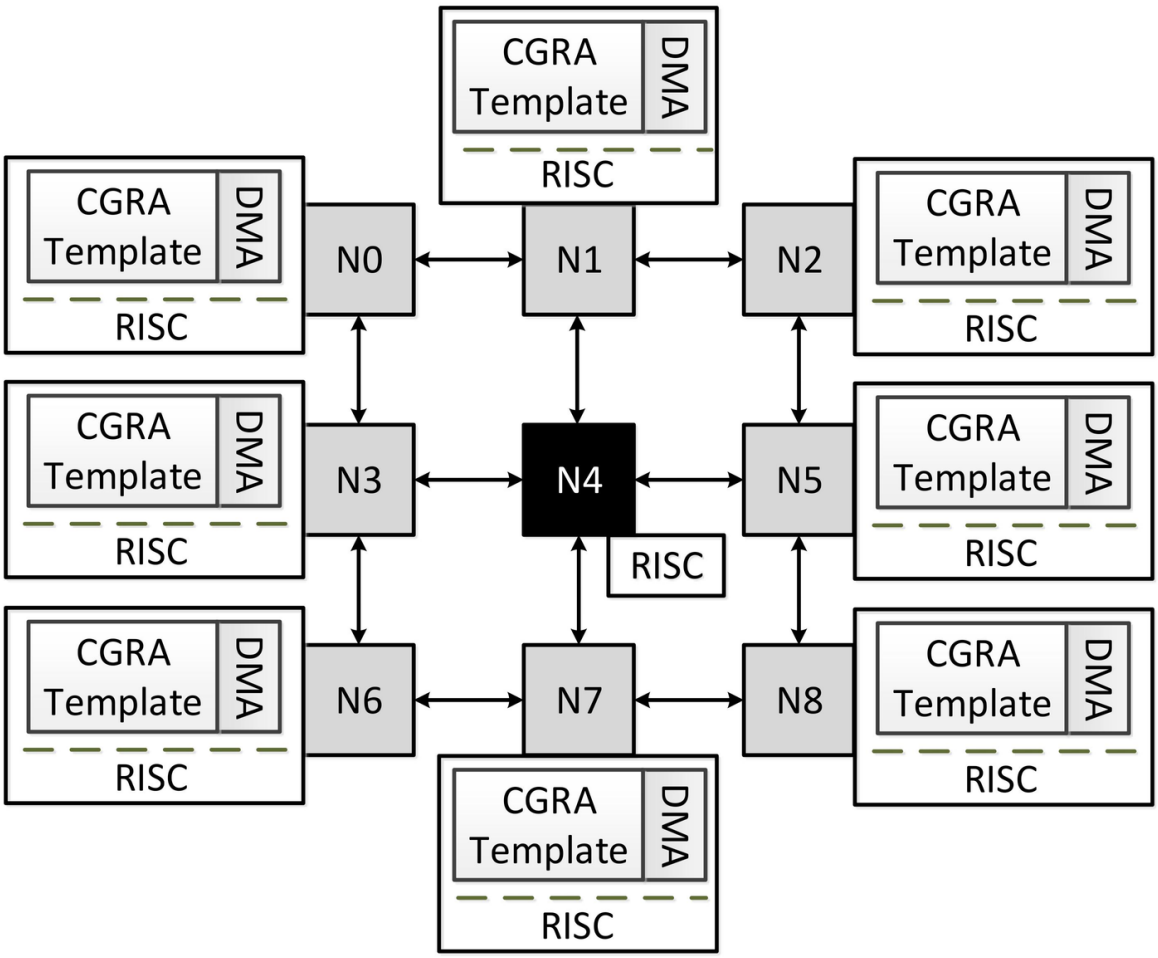

Figure 3.2 *Bridged general view 4/8-point Inverse Discrete Cosine Transform and 4-point Integrated Discreet Sine Transform on HARP © 2018 IEEE [71]*

4. DESIGN AND IMPLEMENTATION OF 4/8 POINT IDCT AND 4-POINT IDST ON TEMPLATE-BASED CGRAS

4.1 4-Point IDCT

The HEVC standard requires core transform matrices of different sizes such as 4×4, 8×8, and 16×16 for the two-dimensional transform. It requires the two-dimensional transform, which resembles to IDCT for all type of transform sizes [69]. In order to carry out the transform, there are various principles and elements, which needs to be considered for carrying out the transform. Either it is a 4 point IDCT or 8 point IDCT, every transform is carried out under specific principles. First, if the use of transform is considered in case of video coding, it can be analyzed that it is implemented to the residual signal, which is retrieved from the intra or inter-frame predictions. The residual signal at the encoder is divided into the square blocks having size N×N where $N = 2^M$ and M is the integer [69]. After that, each residual block is then inputted to the two dimensional transform. The two-dimensional transform can be further implemented as a one dimensional transform to every row and column separately and then the resulting N×N transform coefficients can be further subjected to the quantization in order to gain the quantized transform coefficients [69]. The quantized transform coefficients are then further de-quantized at the decoder and a separate inverse transform is implemented to the gained de-quantized transform coefficients, which result in the outstanding block of the quantized samples. These are then further added to the intra or inter-prediction samples for gaining the reconstructed block [69]. In the case of HEVC, the de-quantized process is specified while the quantization process and ward transforms are further selected by the implementer.

For carrying out the direct cosine transform, the N transform coefficients w_i of the N-Point 1-D DCT is applied to the input sample and u_i can be expressed as follows:

$$w_i = \sum_{j=0}^{N-1} u_j C_{ij} \tag{4.1}$$

while the value of $i = 0, , N-1$. If the elements of C_{ij} of the direct cosine transform matrix C are considered, it can be defined as

$$C_{ij} = \frac{A}{\sqrt{N}} cos\Big[\frac{\pi}{N}\Big(j + \frac{1}{2}\Big)i\Big] \tag{4.2}$$

Here, $i, j = 0, , N-1$ and the value of A is 1 and $2^{\frac{1}{2}}$ for $i = 0; i > 0$ respectively. Additionally, C_i ,which is the basis vector of the DCT can be defined as $C_i = [C_{i0}, , C_{i(N-1)}]^T$ and the value of $i = 0, , N-1$

The direct cosine transform have various properties, which are not only useful for compression but also for the efficient implementation. There are various properties such as ability of basis vectors to provide good energy compaction and their ability to simplify the quantization process [69].If the inverse direct cosine transform is considered, it can be analyzed that with simple matrix multiplications, the number of operations required for 1-D inverse transform is $N(N-1)$ additions and N^2 multiplications. While in case of 2-D transform, the number of additions and multiplications required are $2N^2(N-1)$ and $2N^3$ respectively [69]. However, with the utilization of the symmetry properties of every basis vector retrieved from the IDCT, the arithmetic operations can be reduced. The algorithm is referred as the partial butterfly or Even-Odd decomposition [69]. Now, if the case of 4-point IDCT is considered, it can be analyzed that the unique symmetry properties D_4 is equal to:

$$D_4 = \begin{bmatrix} d^{32}_{16,0} & d^{32}_{16,0} & d^{32}_{16,0} & d^{32}_{16,0} \\ d^{32}_{8,0} & d^{32}_{24,0} & -d^{32}_{24,0} & -d^{32}_{8,0} \\ d^{32}_{16,0} & -d^{32}_{16,0} & -d^{32}_{16,0} & d^{32}_{16,0} \\ d^{32}_{24,0} & -d^{32}_{8,0} & d^{32}_{8,0} & d^{32}_{24,0} \end{bmatrix} \tag{4.3}$$

In order to simplify the notion, the constants $d^{32}_{16,0}$ will be replaced with the d_i. According to the new notion, the inverse transform matrix will be as follows:

$$D_4 = \begin{bmatrix} d_{16} & d_{16} & d_{16} & d_{16} \\ d_8 & d_{24} & -d_{24} & -d_8 \\ d_{16} & -d_{16} & -d_{16} & d_{16} \\ d_{24} & -d_8 & d_8 & d_{24} \end{bmatrix} \tag{4.4}$$

The inverse transform matrix is provided by D_4^T. Assume the value of $x = [x_0, x_1, x_2, x_3]^T$ and it is the input vector while the value of $y = [y_0, y_1, y_2, y_3]^T$ and it denotes the output. According to above provided formula, the 1-D 4-point inverse transform is as follows:
The Even-Odd decomposition of the inverse transform of the N-point input will be carried out through following three steps:

1. Calculation of the even part with the utilization of the $\frac{N}{2} \times \frac{N}{2}$ subset matrix retrieved from the even columns of the inverse transform matrix [69].
2. Calculation of the odd parts with the utilization of the $\frac{N}{2} \times \frac{N}{2}$ subset matrix retrieved from the odd columns of the inverse transform matrix [69].
3. Addition and subtraction of the odd and even parts for generating the N-point output [69].

The decomposition of the inverse 4-point transform is given as follows: Even part:

$$\begin{bmatrix} z_0 \\ \\ z_1 \end{bmatrix} = \begin{bmatrix} d_{16} & d_{16} \\ \\ d_{16} & -d_{16} \end{bmatrix} \times \begin{bmatrix} x_0 \\ \\ x_2 \end{bmatrix} \tag{4.5}$$

The even part can be simplified further as follows:

$$\begin{aligned} t_0 &= d_{16}\, x_0 \\ t_1 &= d_{16}\, x_2 \\ \begin{bmatrix} z_0 \\ \\ z_1 \end{bmatrix} &= \begin{bmatrix} t_0 + t_1 \\ \\ t_0 - t_1 \end{bmatrix} \end{aligned} \tag{4.6}$$

For odd Part we have:

$$\begin{bmatrix} z_2 \\ \\ z_3 \end{bmatrix} = \begin{bmatrix} -d_{24} & d_8 \\ \\ -d_8 & -d_{24} \end{bmatrix} \times \begin{bmatrix} x_1 \\ \\ x_3 \end{bmatrix} \tag{4.7}$$

Addition and Subtraction:

$$\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} z_0 - z_3 \\ z_1 - z_2 \\ z_1 + z_2 \\ z_0 + z_3 \end{bmatrix} \tag{4.8}$$

The direct 1-D 4-point transform requires 12 additions and 16 multiplications. While in case of 2-D transform, it would require 128 multiplications and 96 additions [69]. On the other end, the Even-odd decomposition would require 8 additions and 6 multiplications for 1-D transform [69]. In case of 2-D transform, the Even-Odd decomposition would require 64 additions and 48 multiplications, which saves much than the direct matrix multiplication [69]. Generally, the number of multiplications and additions required for carrying out the IDCT can be found out by implementing the following formula:

$$O_{multiplication} = 2N\left(1 + \sum_{k=1}^{log_2 N} 2^{2k-2}\right) \tag{4.9}$$

$$O_{addition} = 2N\left(\sum_{k=1}^{log_2 N} 2^{2k-1}(2^{2k-1} + 1)\right) \tag{4.10}$$

The arithmetic operations in case of inverse transform (IT) can be additionally lessen lessen the assumption of the value of zero-valued input transform coefficients. In case of HEVC decoder, the information about the input transform is obtained from the de-quantization process.

Now, If the designing of the 4-Piont IDCT is considered in the proposed project, it can be analyzed that the IDCT is primarily the expression of the sequences of limited data points along with the attachment to the addition of cosine functions, which are operating at different frequencies. For developing the 4-point Inverse Direct Cosine Transform (IDCT) accelerator according to the template, which is centered on CGRA, it can be accessed easily from Equation 4.11 and Equation 4.12 in such a manner, which resemble to the creation of the butterfly. If the kernel is considered, it can be analyzed that it is mapped on template-based CGRA and it is the same as where the processing of data occurs during the second context. One

of the important fact, which is interesting is that the first context of every design is carried out to insert the values in it and these values are equal to 12. These values are inserted into the Processing Elements, which are further used in carrying out the shift operation after every time when multiplication occurs. It is also required to help for the prevention of the overflow of data at every time when multiplication is completed along with the fact that every number is represented in 12 bits order to maintaining the accuracy at such a level, which is acceptable. The 4-point IDCT has the matrix coefficients, which have the transform matrix coefficient, which efficient for the 4 by 4 IDCT along with the values of parameters as a=64, b=83, and c=63. The simplification of the equation according to the above-mentioned matrix will produce the output as shown below in the figure in, which 6 multiplications, 5 additions, and 1 subtraction has been carried out. The first and second equation for DCT are as follows:

$$IDCT_{TCoeff}\ 4 \times 4 = \begin{bmatrix} a & c & a & b \\ a & b & -a & -c \\ a & -b & -a & c \\ a & -c & a & b \end{bmatrix} \tag{4.11}$$

$$\begin{aligned} Y_0 &= a(X_0 + X_2) + cX_1 + bX_3 \\ Y_1 &= a(X_0 - X_2) + bX_1 - cX_3 \\ Y_2 &= a(X_0 - X_2) - bX_1 + cX_3 \\ Y_3 &= a(X_0 + X_2) - cX_1 + bX_3 \end{aligned} \tag{4.12}$$

The diagram for Second Contexts in order to carry out the calculation of 4-point IDCT is as follows:

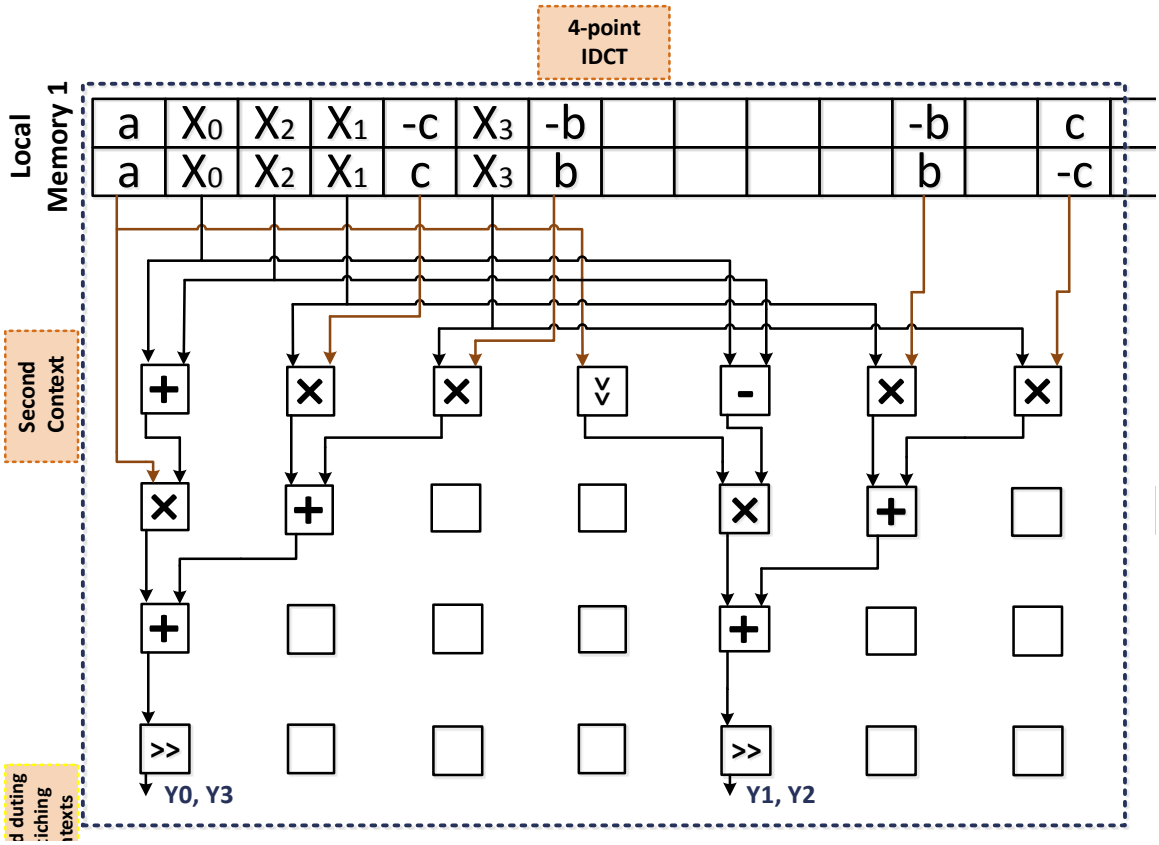

Figure 4.1 *Second Context for the Calculation of 4-point IDCT © 2018 IEEE [71]*

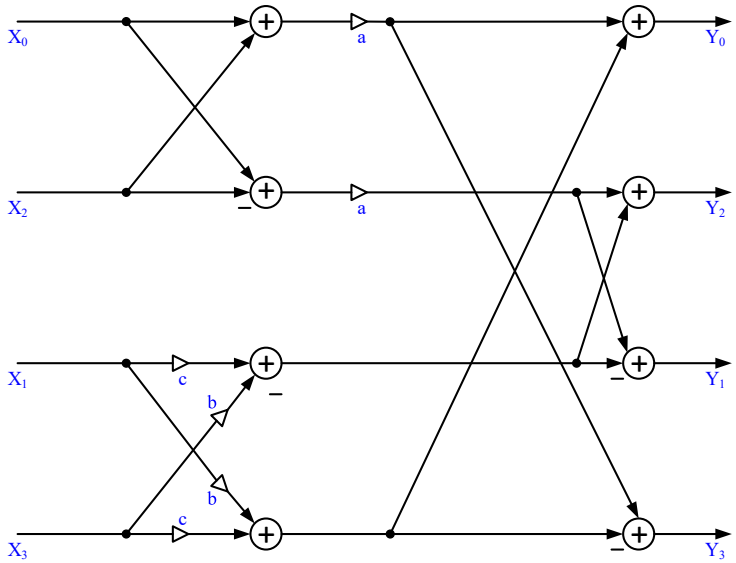

Figure 4.2 *: Butterfly Diagram for 4-Point IDCT*

4.2 8-Point IDCT

8-Point IDCT also follows the same principles and transform is carried out in the same manner as it is carried out in case of 4-Point IDCT. The 8-Point 1-D inverse transform is carried out with the provided formula:

$$y = D_8^T x$$

Where $x = [x_0, x_1, \ldots, x_7]^T$ is the input and $y = [y_0, y_1, \ldots, y_7]^T$ is the output and the transform D_8 is provided by as follows:

$$D_8 = \begin{bmatrix} d_{16} & d_{16} & d_{16} & d_{16} & d_{16} & d_{16} & d_{16} & d_{16} \\ d_4 & d_{12} & d_{20} & d_{28} & -d_{28} & -d_{20} & -d_{12} & -d_4 \\ d_8 & d_{24} & -d_{24} & -d_8 & -d_8 & -d_{24} & d_{24} & d_8 \\ d_{12} & -d_{28} & -d_4 & -d_{20} & d_{20} & d_4 & d_{28} & -d_{12} \\ d_{16} & -d_{16} & -d_{16} & d_{16} & d_{16} & -d_{16} & -d_{16} & d_{16} \\ d_{20} & -d_4 & d_{28} & d_{12} & -d_{12} & -d_{28} & d_4 & -d_{20} \\ d_{24} & -d_8 & d_8 & -d_{24} & -d_{24} & d_8 & -d_8 & -d_{24} \\ d_{28} & -d_{20} & d_{12} & -d_4 & d_4 & -d_{12} & d_{20} & -d_{28} \end{bmatrix} \tag{4.13}$$

In case of 8-point IDCT transform, same rules are applied as applied in case of 4-point IDCT and the decomposition of matrix is carried out to simplify it and to decrease arithmetic processes. The Even-Odd differentiation for the 8-point inverse transform is as follow:

Even Part:

$$\begin{bmatrix} z_0 \\ z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} d_{16} & d_8 & d_{16} & d_{24} \\ d_{16} & d_{24} & -d_{16} & -d_8 \\ d_{16} & -d_{24} & -d_{16} & d_8 \\ d_{16} & -d_8 & d_{16} & -d_{24} \end{bmatrix} \times \begin{bmatrix} x_0 \\ x_2 \\ x_4 \\ x_6 \end{bmatrix} \tag{4.14}$$

Odd Part:

$$\begin{bmatrix} z_4 \\ z_5 \\ z_6 \\ z_7 \end{bmatrix} = \begin{bmatrix} -d_{28} & d_{20} & -d_{12} & d_4 \\ -d_{20} & d_4 & -d_{28} & -d_{12} \\ -d_{12} & -d_{28} & -d_4 & d_{20} \\ -d_4 & -d_{12} & -d_{20} & -d_{28} \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_3 \\ x_5 \\ x_7 \end{bmatrix} \tag{4.15}$$

Addition/subtraction:

$$y = [z_0 - z_7, z_1 - z_6, z_2 - z_5, z_3 - z_4, z_3 + z_4, z_2 + z_5 z_1 + z_6, z_0 + z_7] \tag{4.16}$$

Another point, which is worth mentioning here is that the even part of the 8-point IDCT is the 4-point inverse transform i.e.,

$$\begin{bmatrix} z_0 \\ z_1 \\ z_2 \\ z_3 \end{bmatrix} = D4^T \begin{bmatrix} x_0 \\ x_2 \\ x_4 \\ x_6 \end{bmatrix} \tag{4.17}$$

So, it can be analyzed that the Even decomposition of the 4-point inverse transform can be further used for reducing the computational complexity of the even part [69]. If the multiplications and additions part is considered, the direct 1-D 8 point transform would require 56 additions and 64 multiplications. While in case of 2-D transform, there would be 896 additions and 1024 multiplications will be carried out [69]. In case of Even-Odd decomposition, there would be required 22 multiplications and 28 additions. While in case of 2-D transform with the use of butterfly approach, 448 additions and 352 multiplications will be required. The butterfly diagram for 1-D 8-point IDCT is as follows:

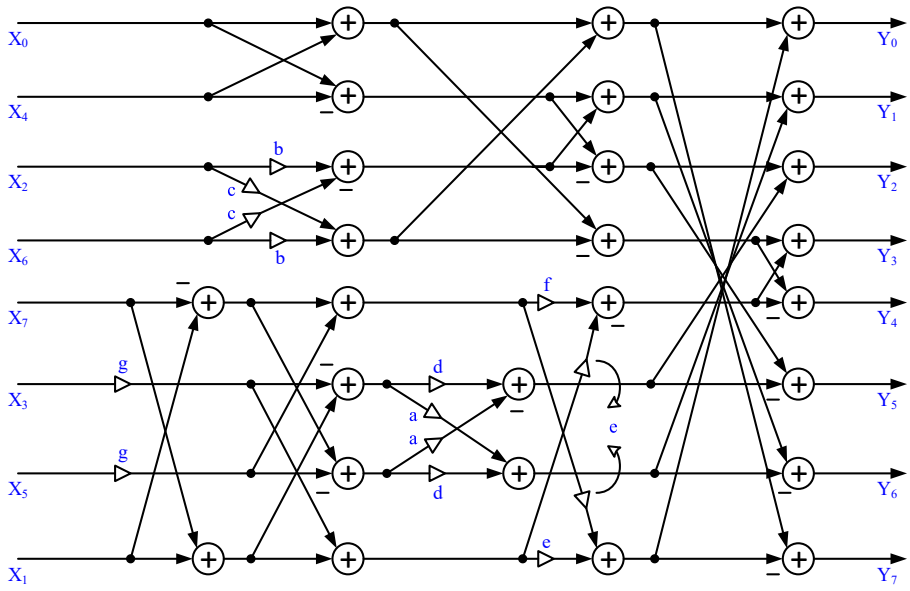

Figure 4.3 *: Butterfly Diagram for 8-Point IDCT*

As discussed above, same logic has been applied to carry out the designing of the 8-point IDCT by the template based on the CGRA's, a very simple process, which includes matrices and basic computation along with mapping. 8-point IDCT co-efficient matrices have been able to be shown with specific regard to the matrices and as a result of the symmetry, which has been factorized in a specific manner such as the successive computations [71]. Moreover, it has only involved the even coefficients, which have been indexed separately and has not been included in the computations. Additionally, only those have been characterized by the coefficients, which have not been indexed i.e., odd indexed. As shown in the calculations, the values of the parameters, which has been used were similar to core transform design in the high-efficiency video coding. For assisting the effective and efficient mapping of the matrices on the template based on CGRA, there is a requirement for them to be simple and divided into a little more in order to ensure that the arithmetic resources and their placement in every PE is calculated in such a way that it is not only efficient but it is also appropriate for the above-mentioned case. After the completion of this case, the transportation of the cases will be carried out in the same way as it was carried out before because the development of the butterfly has begun from the input coefficients with an aim of reaching the output coefficients. As shown in the above figure, two contexts has been used during the mapping exercise

for simplified equations. Additionally, a total of 15 additions, 16 multiplications, and 3 subtractions has been carried out.

$$
\begin{aligned}
Y_0 &= a\underbrace{(X_0 + X_4)}_{Z00} + \underbrace{(bX_2 + cX_6)}_{Z01} + \underbrace{(dX_1 + eX_3 + fX_5 + gX_7)}_{Z02} \\
Y_1 &= a\underbrace{(X_0 - X_4)}_{Z10} + \underbrace{(cX_2 - bX_6)}_{Z11} + \underbrace{(eX_1 - gX_3 - dX_5 - fX_7)}_{Z12} \\
Y_2 &= a\underbrace{(X_0 - X_4)}_{Z20} - \underbrace{(cX_2 - bX_6)}_{Z21} + \underbrace{(fX_1 - dX_3 + gX_5 + eX_7)}_{Z22} \\
Y_3 &= a\underbrace{(X_0 + X_4)}_{Z30} - \underbrace{(bX_2 + cX_6)}_{Z31} + \underbrace{(gX_1 - fX_3 + eX_5 - dX_7)}_{Z32} \\
Y_4 &= a\underbrace{(X_0 + X_4)}_{Z30} - \underbrace{(bX_2 + cX_6)}_{Z31} - \underbrace{(gX_1 - fX_3 + eX_5 - dX_7)}_{Z32} \\
Y_5 &= a\underbrace{(X_0 - X_4)}_{Z20} - \underbrace{(cX_2 - bX_6)}_{Z21} - \underbrace{(fX_1 - dX_3 + gX_5 + eX_7)}_{Z22} \\
Y_6 &= a\underbrace{(X_0 - X_4)}_{Z10} + \underbrace{(cX_2 - bX_6)}_{Z11} - \underbrace{(eX_1 - gX_3 - dX_5 - fX_7)}_{Z12} \\
Y_7 &= a\underbrace{(X_0 + X_4)}_{Z00} + \underbrace{(bX_2 + cX_6)}_{Z01} - \underbrace{(dX_1 + eX_3 + fX_5 + gX_7)}_{Z02}
\end{aligned}
\tag{4.18}
$$

The diagram for 8-point IDCT is as follows

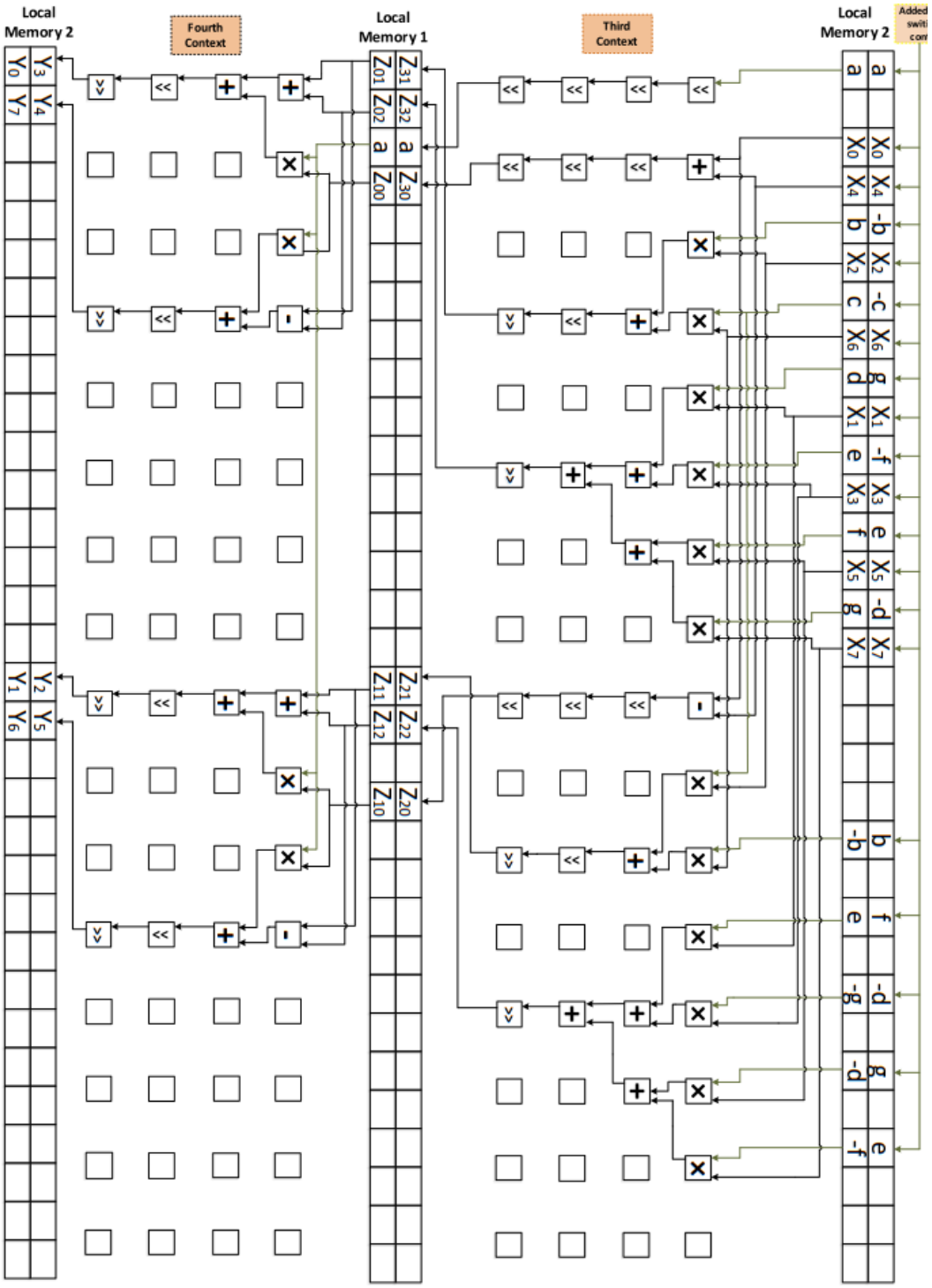

Figure 4.4 *Third & Fourth Contexts for the Calculation of 8-point IDCT © 2018 IEEE [71]*

4.3 4-Point IDST

The Discrete Sine Transform is considered as one of the unitary transforms, which is used for exemplifying and symbolizing the certain signals. If it is used in case of mapping then the two dimensional image can be further exemplified by the matrix, which can itself be expandable known as basis images, which is produced by the unitary matrices [70]. If there is an image, which is exemplified by $N \times N$ matrix, it is quite possible to visualize it as the $N^2 \times 1$ vector [70]. In order to understand the IDST in a better manner, consider the IDST computation of the discrete frequency signal as

$$x(m) = \{1, 2, 3\} = \sigma(m) + 2\sigma(m-1) + 3\sigma(m-2) \tag{4.19}$$

In this case, the length of the sequence is $N = 3$. If the Inverse Discrete Sine Transform is taken on both sides of the above mentioned equation along with the usage of the definition of the IDST, which is as follows [70]:

$$x(n) = \sqrt{\frac{2}{N}} \sum_{m=0} N - 1 X_{VI}^{S}(m) \sin \frac{(2m+1) + (2n+1)\pi}{4N} \tag{4.20}$$

The result obtained is as follows:

$$x(n) = \sqrt{\frac{2}{3}} \sum_{m=0} 2X(m) \sin \frac{(2m+1) + (2n+1)\pi}{12} \tag{4.21}$$

While $0n, m2$. With the simplification of the Equation 4.21 , we get the result as follows:

$$x(n) = 0.8165[\sin \frac{(2n+1)\pi}{12} + 2\sin \frac{(2n+1)\pi}{4} + 3\sin \frac{(2n+1)5\pi}{12}] \tag{4.22}$$

By putting the values of n, the results obtained as follows:

$$x(0) = 0.8165[\sin \frac{\pi}{12} + 2\sin \frac{\pi}{4} + 3\sin \frac{5\pi}{12}] = 3.732$$

$$x(1) = 0.8165[\sin \frac{\pi}{4} + 2\sin \frac{3\pi}{4} + 3\sin \frac{5\pi}{12}] = 0$$

$$x(2) = 0.8165[\sin \frac{5\pi}{12} + 2\sin \frac{5\pi}{4} + 3\sin \frac{25\pi}{12}] = 0.2679$$

The IDST can also be computed using another approach, which is in terms of matrix. The proposed project has also implemented the same approach in order to carry out the arithmetic operations. In order to carry out the -IDST with the help of matrix, the IDST can be found as follows [70]:

$$\begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_{(N-1)} \end{bmatrix} = \sqrt{\frac{2}{N}} \begin{bmatrix} \sin\frac{\pi}{4N} & \dots & \sin\frac{(2N-1)\pi}{4N} \\ \sin\frac{3\pi}{4N} & \dots & \sin\frac{(2N-1)3\pi}{4N} \\ \sin\frac{5\pi}{4N} & \dots & \sin\frac{(2N-1)5\pi}{4N} \\ & & \vdots \\ \sin\frac{\pi}{4N} & \dots & \sin\frac{(2N-1)^2\pi}{4N} \end{bmatrix} \begin{bmatrix} X_0 \\ X_1 \\ X_2 \\ \vdots \\ X_{(N-1)} \end{bmatrix} \tag{4.23}$$

So, if the values are put into the matrix for finding out the IDST, the result obtained is as follows:

$$\begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix} = \sqrt{\frac{2}{3}} \begin{bmatrix} \sin\frac{\pi}{12} & \sin\frac{3\pi}{12} & \sin\frac{5\pi}{12} \\ \sin\frac{3\pi}{12} & \sin\frac{9\pi}{12} & \sin\frac{15\pi}{12} \\ \sin\frac{5\pi}{12} & \sin\frac{15\pi}{12} & \sin\frac{25\pi}{12} \end{bmatrix} \begin{bmatrix} X_0 \\ X_1 \\ X_2 \end{bmatrix} \tag{4.24}$$

The transform coefficient matrix of 4×4 IDST had the matching as compared to the 4×4 IDCT with the values of parameter a=0.1379, b=0.3928, c=0.5879, and d=0.6935 [70]. The matrix was found to be 4 by 4 with 16 variables. Another interesting fact is that the coefficient of DST 4×4 was found to be adding and multiplying values together at the same time. To get the results, there were 4 equations, which can be retrieved and mapped on the CGRA in single context. It can be analyzed from the figure that 6 additions and 8 multiplications were carried out in order to simplify the equation. The equations, which has been used are as follows:

$$\begin{aligned} Y_0 &= aX_0 + bX_1 + cX_2 + dX_3 \\ Y_1 &= bX_0 + dX_1 + aX_2 - cX_3 \\ Y_2 &= cX_0 + aX_1 - dX_2 + bX_3 \\ Y_3 &= dX_0 - cX_1 + bX_2 - aX_3 \end{aligned} \tag{4.25}$$

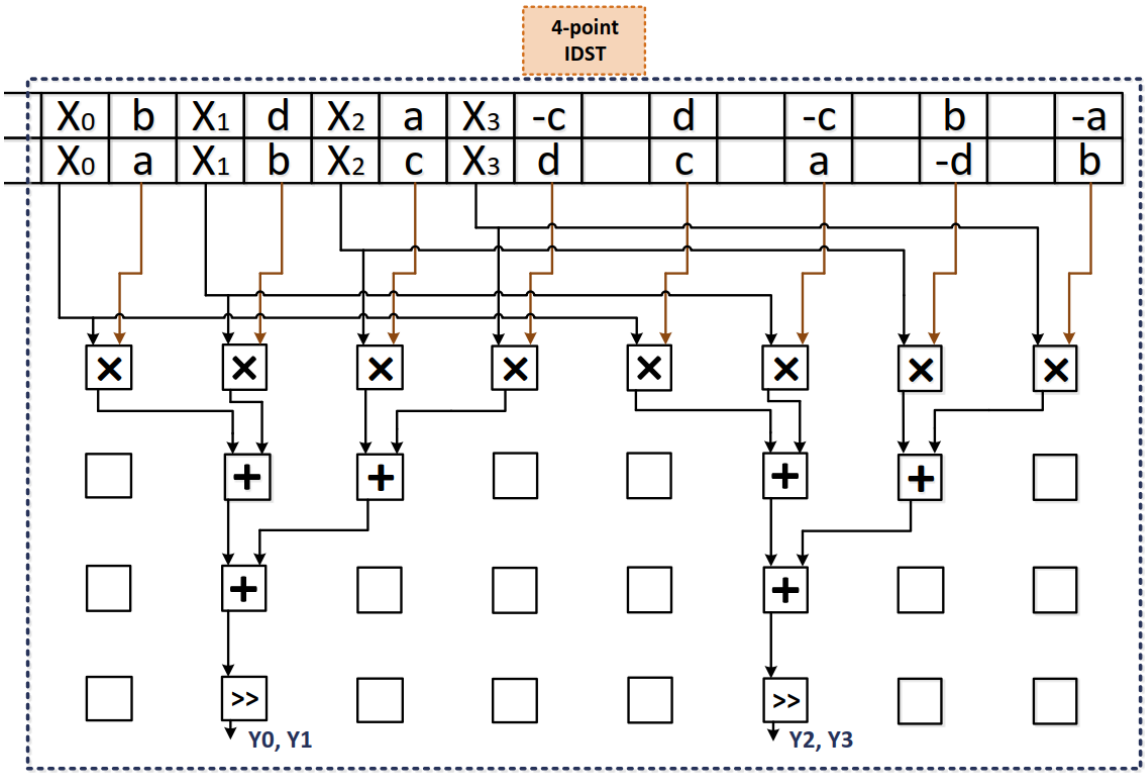

Figure 4.5 *Second Contexts for 4-point IDST © 2018 IEEE [71]*

Figure 4.6 shows signal flow of 4 point IDST according to Equation 4.25

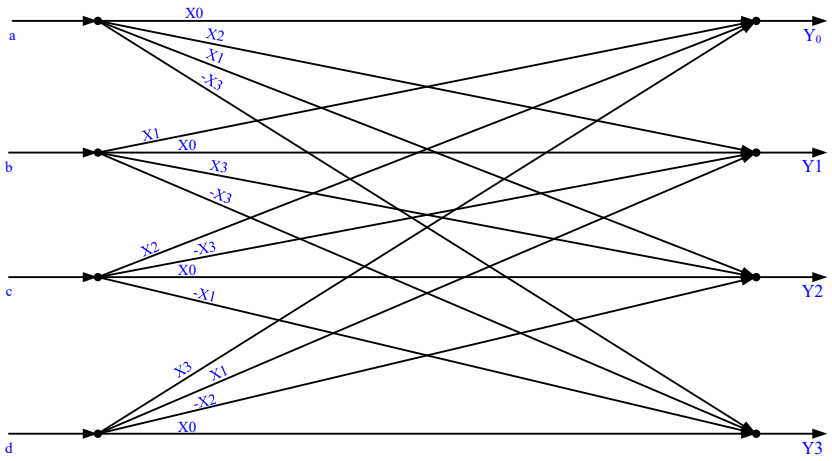

Figure 4.6 *: Signal Flow for 4-Point IDST*

4.4 Implementation of 4/8 Point IDST/IDCT ACCELERATOR ON HARP

It can be seen from the Figure 4.7 that there are three RISC cores, which are integrated into the middle row of the HARP template and N4 is acting as the supervisor node. While in the meantime, the N3 and N5 RISC cores have the responsibility of controlling the associated CGRA nodes. Following the loading of configuration streams by the CGRA nodes, the data that has to be processed should be loaded into local memories in a parallel manner. According to the devised target set by the developer, the first, second, and third context can be empowered for performing the 4-point IDCT/IDST or 8-point IDCT respectively. The total number of clock cycles, which are required for transferring the data for processing along with the transpose matrix coefficients for 8-point IDCT and 4-point IDCT/IDST is 2456 CC. N-point 1-D transform can be implemented to every column and row for carrying out the implementation of the 2-D transform operations.
It can be understood with an assumption that 4-point 2-D IDCT/IDST is going to be processed with the template based CGRAs and if each row and column of the matrix is considered as a stream, there are total 8 streams, which should be loaded into the local memories of the CGRA nodes with the fact that each CGRA node is allocated with 2 streams. While all the CGRA nodes are running and the second context is enabled. Therefore, a 2-D 4-point IDCT/IDST and a 1-D 4-point IDCT/IDST can be implemented in 121 CC and 56 CC respectively. In case of 8-point 2-D IDCT, 16 streams should be inserted sequentially into the system and 4 streams should be allotted to each CGRA node while the third and fourth contexts are enabled, the 2-D 8-point and 1-D 8-point IDCT can be implemented in 182 CC and 64 CC respectively. With the completion of all executions of rows and columns of 2-D IDST/IDCT, the DMA device will deliver the results from the nodes of CGRA to the data memory of the host RISC processor and assemble them in the data memory of the supervisory node in order to carry out the further processing.

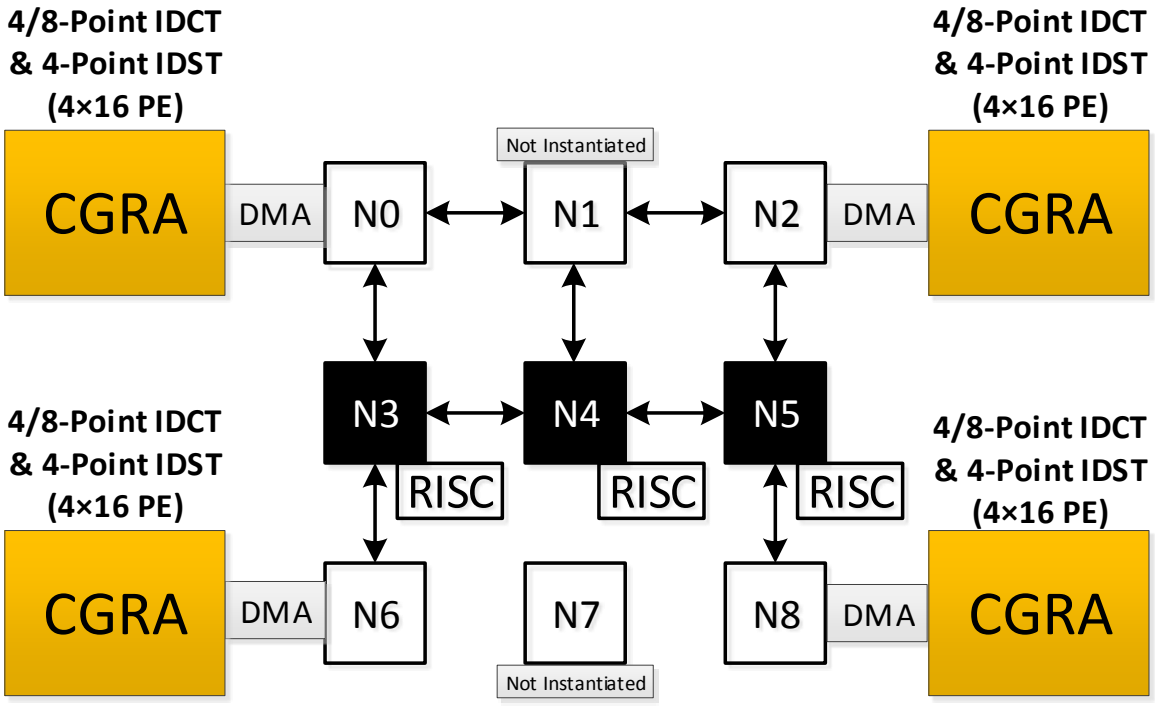

Figure 4.7 *Abridged general view 4/8-Point IDCT and 4-Point IDST on HARP © 2018 IEEE [71]*

5. MEASUREMENTS, ESTIMATIONS, EVALUATION AND COMPARISONS

The FPGA unit Stratix-V (5SGXMB9R3H4C2) has been used for implementation of the hypothesized architecture. There are many different variables used in the testing process. Each variable is set on a certain value (according to researches) and the test is repeated for every trial. There has been two different models used in this study, one model is the fast timing while the other model is the slow timing model (both models were set to 900 mV). The set variable in this trial was the temperature, which was set the first time for 0°C and 85°C. The frequency for the operation was then measured, which was 162.28 MHz for the low temperature while it was 159.84 MHz for the higher temperature. On the other hand, the operating frequencies achieved were 260.42 MHz and 240.62 MHz respective to the temperatures. The frequency chosen for the experiments was an intermediate value of 200 MHz, which to be used for all units in the system studied. 5.1 shows all the items used by an explanation of each node in the system. For the Adaptive Logic Modules (ALMs), a total of 36.9% of it is used for the designed architecture to work. In addition, 276 18 bit DSP units are used for carrying out the 32-bit multiplication.

Table 5.1 *Node-by-node Breakdown of Resource Utilization. © 2018 IEEE [71]*

Node	**ALMs**	**Registers**	**Memory Bits**	**(32-bit Multipliers) DSPs**
N0	24,393	7,862	2,633,472	(30) 60
N2	24,360	7,271	2,633,472	(30) 60
N3	5,588	5,688	3,145,728	(6) 12
N4	5,623	5,732	4,194,304	(6) 12
N5	5,563	5,595	3,145,728	(6) 12
N6	24,368	7,564	2,633,472	(30) 60
N8	24,372	7,889	2,633,472	(30) 60
NoC	2,603	4,207	-	-
Total	116,870 36.9%	51808 8.2%	21,019,648 38.9%	(138) 276 78.4%

On the other hand, the Table 5.2 focuses on the description of dynamic power dissipation through the experiment. This power dissipation lost is measured according to placement and routing (post P&R), of the data using a PowerPlay Power Analyzer Tool of Quartus II 15.0. The variables set for this experiment were the temperature at 25°C (ambient temperature) and the frequency as stated earlier at 200 MHz. At these ratings, the value for the static power lost was 1621.15 mW, as long as the system is ON and working. However, when measuring the dynamic power loss, the value goes up to 2355.2 mW. This increase is due to the signal transition of the data while the IDCT/IDST was used. Another form of power that has been considered is the input thermal power, which when added, gives a total power loss of 4034.16 mW (4.03 W). In addition, Table 5.2 shows the energy used up by every node of the system, which can be calculated by multiplying the power loss with the time for execution. Generally, each CRGA has to be initialized a task at the point of designing (before implementation). This is where the architect will set each node of CGRA to work as a 4-point IDCT/IDST or 8-point IDCT. The difference in this choice will be in the active time, which is showed in 5.2 (divided by a /). For the dynamic power used at each node of CGRA, it is calculated relating to the time taken for the 4-point IDCT/IDST and the 8-point IDCT to work, which was 0.2 μs and 0.32 μs respectively. Thus, the value for the dynamic power lost was 0.1 μJ for the 4-point and 0.12 μJ for the 8-point. Hence, when the second design for the CGRA, which is temple based was added to the system the energy was studied again to observe the difference. For the 4-point IDCT/IDST, the value for the dynamic energy used was 1.76 μJ. On the other hand, for the 8-point IDCT, it would need to use the third and the fourth design architecture. When observing the energy used it showed a value of 3.06 μJ.

Table 5.2 *Dynamic power and energy estimation of each CGRA node and the NoC. GPP and IL stand for General Purpose Processing and Integration Logic, respectively.* *[71]*

Node	Accelerator Type	Dynamic Power (mW)	Active Time (μs)	Dynamic Energy (μJ)
N0	CTX.2/ CTX.3,4	372.51	0.28/0.32	0.1/0.12
N2	CTX.2/ CTX.3,4	371.92	0.28/0.32	0.1/0.12
N3	GPP Synchronization, Control	115.07	5.95/11.17	0.68/1.29
N4		115.02	0	0
N5		115.11	5.95/11.17	0.68/1.29
N6	CTX.2/ CTX.3,4	372.37	0.28/0.32	0.1/0.12
N8	CTX.2/ CTX.3,4	369.19	0.28/0.32	0.1/0.12
NoC	-	10.12	-	-
IL	-	513.87	-	-
Total	-	2355.2	-	1.76/3.06

The HARP used in this design was modified in a ways, which had 256 instantiated Processing Elements (PEs). Now it is important to study the operations per second that took place in the system. Such value is huge as the system itself carries out millions of operations, thus the unit for the value is Giga Operations Per Second (GOPS). To study this factor, the frequency of the system, as stated earlier, is still at 200 MHz with the power loss at 4.03W as observed earlier. At these values, the hypothesized architecture showed a response of 51.2 GOPS and 0.012 GOPS/mW.

For a Full HD 1080p encoding at 30 fps, the minimum requirement for throughput can be known by 1920 x 1080 x 30 / (8×8), which gives a value of 972,000 blocks/second; for the case of the 8-point IDCT. Currently, the value for the HARP system used here can implement the 8-point IDCT but requires a frequency, which is calculated as follows: $972,000 \times 182 = 176.9$ million cycles/second or 176.9 MHz. When looking at the maximum frequency that can be reached, which is the 260.42 MHz, the obtained frequency is less than the maximum, which suggests that the 1080p format at 30 fps is accepted and can be achieved.

6. CONCLUSION

This research study considers implementation of IDCT/IDST on HARP template with the utilization of CGRA. In this thesis, we focus on using the architecture of HARP for the new technology. This is achieved through HEVC rules that uses the IDCT/IDST in their second dimensional form. The 56 and 64 clock cycles are used for the one dimensional IDCT/IDST, which uses one or eight point(for IDCT only) respectively. And processing it by the PE array. This will give an output, which is strong with values of 4.03 W, 0.012 GOPS/mW and 200 MHz in 28 nm chip. This is the strong image known commonly as HD 1080p at 30 frames per second. By using the CGRA template there are many different forms of usage, which include the 4 point IDCT, 8 point IDCT and 4 point IDST. These templates can be used as an accelerator for the HARP technology. This system uses parallel computing, which is how it works efficiently with a strong output. It utilizes the ability of having multitasks being done at the same time, while still in an accurate state. The HARP architecture has 9 nodes, with the node at the center used to control the rest of the system. The difference between implementing the 4 point and the 8 point IDCT focuses on using different matrices giving different sizes for the templates. However, for the 4/8 point IDCT/IDST for the GCRA as an accelerator, the three cores (RISC) are used together with each one having its own supervisor node. According to retrieved results, the video quality result obtained was higher and showed significant improvement as compared to other results obtained in earlier research studies. Even with the few problems faced in this system, in terms of power and data allocation, the technology has reached a better video image.

Currently this system experience power loss issue due to the utilization of multiple kernels in the program. This is something that has to be discussed for future references as power is a very crucial element in the system. This can be done by setting less energy and power towards the core having more controlled values of the voltage and frequency sent to the core. Another problem is the compatibility of the IDCT/IDST with the hardware, which creates a lot of problems due to the language barrier with the designers. However, this approach helps all developers and be able to work effectively without problems. Another add-on that should be done the CGRA should be re-designed in order to reach a better output. It has

been proven that the size, in terms of dimensions, of the CGRA affects the output this it should be changed to a better size.

6.1 Future Work

There has been a very observable use in power and energy which is a great setback for this development. An awareness for the energy problem should be raised as it is an important factor in deciding whether the technology is effective or not. Furthermore, the study of HARP technology and using its architect in image processing is important for future researches as it has shown a great impact in the image field. Apart from that, future work could be done in designing the 16/32 point IDCT and DCT on HARP template in parallel multitasking for using at the time of designing it for other purposes. The integration of 16/32 point IDCT and DCT on HARP template would allow to support and compress larger block sizes. The integration of DCT and IDCT at a parallel manner on HARP would allow to have multi-tasking for CGRA which is not possible in case of executed research study. The current research study implements 4 and 8 point IDCT along with 4 point IDST on HARP. While the extended version of the research would allow to integrate 16/32 point DCT and IDCT in a parallel manner. Another direction which can also be specified for the future work in case of HEVC design on HARP is the employment of FCS and DFS on the specific current design to mitigate the power dissipation issue. As discussed above in the start of the thesis, the power dissipation issue is one of the major issue which makes it difficult in case of compression of higher block sizes. The implementation of FCS and DFS on the current design would help to address the power dissipation. These are two major future directions for current research study carried out in this thesis and it would not only allow to save resources in terms of time and power but would also be a major breakthrough for developing advanced technologies.

BIBLIOGRAPHY

[1] Nouri, S. (2018). Power and Energy Aware Heterogeneous Computing Platform. (Tampere University of Technology. Publication; Vol. 1594). Tampere University of Technology.

[2] Bingfeng Mei, F. -. Veredas and B. Masschelein, "Mapping an H.264/AVC decoder onto the ADRES reconfigurable architecture," International Conference on Field Programmable Logic and Applications, 2005., Tampere, 2005, pp. 622-625. doi: 10.1109/FPL.2005.1515799

[3] H. Singh, Ming-Hau Lee, Guangming Lu, F. J. Kurdahi, N. Bagherzadeh and E. M. Chaves Filho, "MorphoSys: an integrated reconfigurable system for data-parallel and computation-intensive applications," in IEEE Transactions on Computers, vol. 49, no. 5, pp. 465-481, May 2000. doi: 10.1109/12.859540

[4] Lee, MH., Singh, H., Lu, G. et al. The Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology (2000) 24: 147. https://doi.org/10.1023/A:1008189221436

[5] Guangming Lu et al., "The MorphoSys dynamically reconfigurable system-on-chip," Proceedings of the First NASA/DoD Workshop on Evolvable Hardware, Pasadena, CA, USA, 1999, pp. 152-160. doi: 10.1109/EH.1999.785447

[6] B. Mei, S. Vernalde, D. Verkest, H. D. Man, and R. Lauwereins, ADRES: An Architecture with Tightly Coupled VLIW Processor and Coarse-Grained Reconfigurable Matrix., in FPL, 2003, vol. 2778, pp. 6170.

[7] C. B. Ciobanu et al., "EXTRA: Towards an Efficient Open Platform for Reconfigurable High Performance Computing," 2015 IEEE 18th International Conference on Computational Science and Engineering, Porto, 2015, pp. 339-342. doi: 10.1109/CSE.2015.54

[8] H. Singh, Ming-Hau Lee, Guangming Lu, F. J. Kurdahi, N. Bagherzadeh and E. M. C. Filho, "MorphoSys: a reconfigurable architecture for multimedia applications," Proceedings. XI Brazilian Symposium on Integrated Circuit Design (Cat. No.98EX216), Rio de Janeiro, Brazil, 1998, pp. 134-139. doi: 10.1109/SBCCI.1998.715427

[9] V. Baumgarte, G. Ehlers, F. May, A. Naeckel, M. Vorbach, and M. Weinhardt, PACT XPPA Self-Reconfigurable Data Processing Architecture, Journal of Supercomputing, vol. 26, no. 2, pp. 167–184, 2003.

[10] J. M. P. Cardoso and M. Weinhardt, Xpp-vc: A c compiler with temporal partitioning for the pact-xpp architecture, in Field-Programmable Logic and Applications: Reconfigurable Computing Is Going Mainstream, Editors: M. Glesner and P. Zipf and M. Renovell, Lecture Notes in Computer Science, Springer Berlin Heidelberg, vol. 2438, pp. 864874, 2002.

[11] H. Hoffmann et al., "Self-aware computing in the Angstrom processor," DAC Design Automation Conference 2012, San Francisco, CA, 2012, pp. 259-264. doi: 10.1145/2228360.2228409

[12] W. Hussain, R. Airoldi, H. Hoffmann, T. Ahonen and J. Nurmi, "Design of an accelerator-rich architecture by integrating multiple heterogeneous coarse grain reconfigurable arrays over a network-on-chip," 2014 IEEE 25th International Conference on Application-Specific Systems, Architectures and Processors, Zurich, 2014, pp. 131-138. doi: 10.1109/ASAP.2014.6868647

[13] W. Hussain, R. Airoldi, H. Hoffmann, T. Ahonen and J. Nurmi, "HARP2: An X-Scale Reconfigurable Accelerator-Rich Platform for Massively-Parallel Signal Processing Algorithms", Journal of Signal Processing Systems, vol. 85, no. 3, pp. 341-353, 2015. Available: 10.1007/s11265-015-1054-9.

[14] N. Voros et al., "MORPHEUS: A heterogeneous dynamically reconfigurable platform for designing highly complex embedded systems", ACM Transactions on Embedded Computing Systems, vol. 12, no. 3, pp. 1-33, 2013. Available: 10.1145/2442116.2442120.

[15] F. Thoma et al., "MORPHEUS: Heterogeneous Reconfigurable Computing," 2007 International Conference on Field Programmable Logic and Applications, Amsterdam, 2007, pp. 409-414. doi: 10.1109/FPL.2007.4380681

[16] D. Rossi, F. Campi, S. Spolzino, S. Pucillo and R. Guerrieri, "A Heterogeneous Digital Signal Processor for Dynamically Reconfigurable Computing," in IEEE Journal of Solid-State Circuits, vol. 45, no. 8, pp. 1615-1626, Aug. 2010. doi: 10.1109/JSSC.2010.2048149

[17] D. Melpignano et al., "Platform 2012, a many-core computing accelerator for embedded SoCs: Performance evaluation of visual analytics applications," DAC Design Automation Conference 2012, San Francisco, CA, 2012, pp. 1137-1142. doi: 10.1145/2228360.2228568

[18] F. Conti, C. Pilkington, A. Marongiu and L. Benini, "He-P2012: Architectural heterogeneity exploration on a scalable many-core platform," 2014 IEEE 25th

International Conference on Application-Specific Systems, Architectures and Processors, Zurich, 2014, pp. 114-120. doi: 10.1109/ASAP.2014.6868645

[19] F. Garzia, W. Hussain and J. Nurmi, "CREMA: A coarse-grain reconfigurable array with mapping adaptiveness," 2009 International Conference on Field Programmable Logic and Applications, Prague, 2009, pp. 708-712. doi: 10.1109/FPL.2009.5272353

[20] N. Voros, M. Hbner, D. Ghringer, and C. Antonopoulos, Applied reconfigurable computing. architectures, tools, and applications, 14th International Symposium, ARC 2018, Santorini, Greece, May 2-4, 2018, Proceedings, Springer, 2018.

[21] W. Hussain, F. Garzia, T. Ahonen and J. Nurmi, "Designing Fast Fourier Transform Accelerators for Orthogonal Frequency-Division Multiplexing Systems", Journal of Signal Processing Systems, vol. 69, no. 2, pp. 161-171, 2011. Available: 10.1007/s11265-011-0642-6.

[22] M. L. Ferreira, J. C. Ferreira, and M. Hbner, A parallel-pipelined ofdm baseband modulator with dynamic frequency scaling for 5g systems, Applied Reconfigurable Computing. Architectures, Tools, and Applications, Springer International Publishing AG, part of Springer Nature 2018, 2018.

[23] Altera, Altera Product Catalog. 2015 Version 15.0,. www.altera.com: Altera Corporation, 2014.

[24] J. Heiskala and J. Terry, Ofdm wireless lans: A theoretical and practical guide, 2002 by Sams Publishing, SAMS, 201 West 103rd St., Indianapolis, Indiana, 46290 USA

[25] Altera, Design Implementation and Optimization Quartus-II Handbook Version 13.1. Altera Corporation, 2013.

[26] P. D. Hennessy JL, Computer architecture: A quantitative approach. 3rd edn. elseview morgan kaufmann, san francisco. 1990.

[27] C. Panis, Vliw dsp processor for high-end mobile communication applications, In Processor Design: System-on-Chip Computing for ASICs and FPGAs, J. Nurmi, Ed. Kluwer Academic Publishers / Springer Publishers, pp. 83100, 2007.

[28] N. Vassiliadis, N. Kavvadias, G. Theodoridis and S. Nikolaidis, "A RISC architecture extended by an efficient tightly coupled reconfigurable unit", International Journal of Electronics, vol. 93, no. 6, pp. 421-438, 2006. Available: 10.1080/00207210600565127.

[29] F. Garzia, T. Ahonen and J. Nurmi, "A switched interconnection infrastructure to tightly-couple a RISC processor core with a coarse grain reconfigurable array," 2009 Ph.D. Research in Microelectronics and Electronics, Cork, 2009, pp. 16-19. doi: 10.1109/RME.2009.5201372

[30] Xilinx, `http://www.xilinx.com`

[31] Altera, `http://www.altera.com`

[32] J. R. Hauser and J. Wawrzynek, "Garp: a MIPS processor with a reconfigurable co-processor," Proceedings. The 5th Annual IEEE Symposium on Field-Programmable Custom Computing Machines Cat. No.97TB100186), Napa Valley, CA, USA, 1997, pp. 12-21. doi: 10.1109/FPGA.1997.624600

[33] T. J. Callahan, J. R. Hauser and J. Wawrzynek, "The Garp architecture and C compiler," in Computer, vol. 33, no. 4, pp. 62-69, April 2000. doi: 10.1109/2.839323

[34] N. S. Voros, A. Rosti, and M. Hubner, Flexeos embedded fpga solution, in Dynamic System Reconfiguration in Heterogeneous Platforms, Lecture Notes in Electrical Engineering, Springer Netherlands, vol. 40, pp. 3947.

[35] S. Vassiliadis, S. Wong, G. Gaydadjiev, K. Bertels, G. Kuzmanov and E. M. Panainte, "The MOLEN polymorphic processor," in IEEE Transactions on Computers, vol. 53, no. 11, pp. 1363-1375, Nov. 2004. doi: 10.1109/TC.2004.104

[36] C. B. Ciobanu et al., "EXTRA: Towards an Efficient Open Platform for Reconfigurable High Performance Computing," 2015 IEEE 18th International Conference on Computational Science and Engineering, Porto, 2015, pp. 339-342. doi: 10.1109/CSE.2015.54

[37] D. Stroobandt et al., "EXTRA: Towards the exploitation of eXascale technology for reconfigurable architectures," 2016 11th International Symposium on Reconfigurable Communication-centric Systems-on-Chip (ReCoSoC), Tallinn, 2016, pp. 1-7. doi: 10.1109/ReCoSoC.2016.7533896

[38] J. Kylliäinen, T. Ahonen, and J. Nurmi, General-purpose embedded processor cores - the coffee risc example, In Processor Design:System-on-Chip Computing for ASICs and FPGAs, J. Nurmi, Ed. Kluwer Academic Publishers / Springer Publishers, pp. 83 – 100, 2007.

[39] B. Mei, S. Vernalde, D. Verkest, H. De Man, and R. Lauwereins, ADRES: An Architecture with Tightly Coupled VLIW Processor and Coarse-Grained Reconfigurable Matrix, Lecture Notes in Computer Science, pp. 6170, 2003.

[40] M. B. Taylor et al., "The Raw microprocessor: a computational fabric for software circuits and general-purpose programs," in IEEE Micro, vol. 22, no. 2, pp. 25-35, March-April 2002. doi: 10.1109/MM.2002.997877

[41] E. Panainte, K. Bertels and S. Vassiliadis, "The Molen compiler for reconfigurable processors", ACM Transactions on Embedded Computing Systems, vol. 6, no. 1, p. 6, 2007. Available: 10.1145/1210268.1210274.

[42] Z. Chen and Z. Zhang, "A High-Speed 2-D IDCT Processor for Image/Video Decoding," 2009 2nd International Congress on Image and Signal Processing, Tianjin, 2009, pp. 1-4.doi: 10.1109/CISP.2009.5302415

[43] S. Shen, W. Shen, Y. Fan and X. Zeng, "A Unified 4/8/16/32-Point Integer IDCT Architecture for Multiple Video Coding Standards," 2012 IEEE International Conference on Multimedia and Expo, Melbourne, VIC, 2012, pp. 788-793. doi: 10.1109/ICME.2012.7

[44] J. Shan, C. Chen and E. Yang, "High performance 2-D IDCT for Image/Video Decoding based on FPGA," 2012 International Conference on Audio, Language and Image Processing, Shanghai, 2012, pp. 33-38. doi: 10.1109/ICALIP.2012.6376582

[45] K. Swaminathan and J. V. Kumar, "Reconfigurable IDCT architecture on FPGA for multiple video standards," 2012 International Conference on Emerging Trends in Science, Engineering and Technology (INCOSET), Tiruchirappalli, 2012, pp. 264-268. doi: 10.1109/INCOSET.2012.6513916

[46] R. Conceio, J. C. Souza, R. Jeske, M. Porto, J. Mattos and L. Agostini, "Hardware design for the 16×16 IDCT of the HEVC video coding standard," 2013 26th Symposium on Integrated Circuits and Systems Design (SBCCI), Curitiba, 2013, pp. 1-6. doi: 10.1109/SBCCI.2013.6644881

[47] L. Hong, W. He, H. Zhu and Z. Mao, "A cost effective 2-D adaptive block size IDCT architecture for HEVC standard," 2013 IEEE 56th International Midwest Symposium on Circuits and Systems (MWSCAS), Columbus, OH, 2013, pp. 1290-1293. doi: 10.1109/MWSCAS.2013.6674891

[48] Tianlong Ma, Cong Liu, Yibo Fan and Xiaoyang Zeng, "A fast 8×8 IDCT algorithm for HEVC," 2013 IEEE 10th International Conference on ASIC, Shenzhen, 2013, pp. 1-4. doi: 10.1109/ASICON.2013.6811848

[49] R. Conceio, J. C. de Souza, R. Jeske, M. Porto, B. Zatt and L. Agostini, "Power efficient and high troughtput multi-size IDCT targeting UHD HEVC decoders,"

2014 IEEE International Symposium on Circuits and Systems (ISCAS), Melbourne VIC, 2014, pp. 1925-1928. doi: 10.1109/ISCAS.2014.6865537

[50] Z. Yao, W. He, L. Hong, G. He and Z. Mao, "Area and throughput efficient IDCT/IDST architecture for HEVC standard," 2014 IEEE International Symposium on Circuits and Systems (ISCAS), Melbourne VIC, 2014, pp. 2511-2514. doi: 10.1109/ISCAS.2014.6865683

[51] J. Nikara, R. Rosendahl, K. Punkka and J. Takala, "Implementation of two-dimensional discrete cosine transform and its inverse," 2004 12th European Signal Processing Conference, Vienna, 2004, pp. 1537-1540.

[52] R. Conceio, . Arajo, M. Porto, B. Zatt and L. Agostini, "Hardware design of fast HEVC 2-D IDCT targeting real-time UHD 4K applications," 2015 IEEE 6th Latin American Symposium on Circuits & Systems (LASCAS), Montevideo, 2015, pp. 1-4. doi: 10.1109/LASCAS.2015.7250473

[53] E. Kalali and I. Hamzaoglu, "FPGA implementations of HEVC Inverse DCT using high-level synthesis," 2015 Conference on Design and Architectures for Signal and Image Processing (DASIP), Krakow, 2015, pp. 1-6. doi: 10.1109/DASIP.2015.7367262

[54] A. Kilany, M. Abdelrasoul, A. Shalaby and M. S. Sayed, "A reconfigurable 2-D IDCT architecture for HEVC encoder/decoder," 2015 27th International Conference on Microelectronics (ICM), Casablanca, 2015, pp. 242-245. doi: 10.1109/ICM.2015.7438033

[55] K. Yao, R. Wang, Z. Wang, W. Wang and W. Gao, "A Fast and Lossless IDCT Design for AVS2 Codec," 2016 IEEE Second International Conference on Multimedia Big Data (BigMM), Taipei, 2016, pp. 241-245. doi: 10.1109/BigMM.2016.60

[56] P. Sjvall, V. Viitamki, J. Vanne and T. D. Hmlinen, "High-level synthesis implementation of HEVC 2-D DCT/DST on FPGA," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 1547-1551. doi: 10.1109/ICASSP.2017.7952416

[57] V. Viitamki, P. Sjvall, J. Vanne and T. D. Hmlinen, "High-level synthesized 2-D IDCT/IDST implementation for HEVC codecs on FPGA," 2017 IEEE International Symposium on Circuits and Systems (ISCAS), Baltimore, MD, 2017, pp. 1-4. doi: 10.1109/ISCAS.2017.8050323

[58] K. K. Singh and D. Pandey, "Implementation of DCT and IDCT based image compression and decompression on FPGA," 2017 International Conference on Inventive Systems and Control (ICISC), Coimbatore, 2017, pp. 1-4. doi: 10.1109/ICISC.2017.8068640

[59] H. Sun, D. Zhou, J. Zhu, S. Kimura and S. Goto, "An area-efficient 4/8/16/32-point inverse DCT architecture for UHDTV HEVC decoder," 2014 IEEE Visual Communications and Image Processing Conference, Valletta, 2014, pp. 197-200. doi: 10.1109/VCIP.2014.7051538

[60] J. Park, W. Nam, S. Han and S. Lee, "2-D Large Inverse Transform (16×16, 16×16) for HEVC (High Efficiency Video Coding)", JSTS:Journal of Semiconductor Technology and Science, vol. 12, no. 2, pp. 203-211, 2012. Available: 10.5573/jsts.2012.12.2.203.

[61] M. Martuza and K. Wahid, "Low Cost Design of a Hybrid Architecture of Integer Inverse DCT for H.264, VC-1, AVS, and HEVC", VLSI Design, vol. 2012, pp. 1-10, 2012. Available: 10.1155/2012/242989.

[62] E. Kalali, E. Ozcan, O. M. Yalcinkaya and I. Hamzaoglu, "A low energy HEVC Inverse DCT hardware," 2013 IEEE Third International Conference on Consumer Electronics Berlin (ICCE-Berlin), Berlin, 2013, pp. 123-124. doi: 10.1109/ICCE-Berlin.2013.6698021

[63] H. SUN, D. ZHOU, P. LIU and S. GOTO, "A Low-Cost VLSI Architecture of Multiple-Size IDCT for H.265/HEVC", IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol. 97, no. 12, pp. 2467-2476, 2014. Available: 10.1587/transfun.e97.a.2467.

[64] H. Liang, H. Weifeng, Z. Hui and M. Zhigang, "A full-pipelined 2-D IDCT/IDST VLSI architecture with adaptive block-size for HEVC standard", IEICE Electronics Express, vol. 10, no. 9, pp. 20130210-20130210, 2013. Available: 10.1587/elex.10.20130210.

[65] M. Tikekar, C. Huang, V. Sze and A. Chandrakasan, "Energy and area-efficient hardware implementation of HEVC inverse transform and dequantization," 2014 IEEE International Conference on Image Processing (ICIP), Paris, 2014, pp. 2100-2104. doi: 10.1109/ICIP.2014.7025421

[66] "Coarse Grain Reconfigurable Arrays — Compiler Microarchitecture Lab", `http://aviral.lab.asu.edu/cgra`

[67] D. Chisnall, "The Dark Silicon Problem and What it Means for CPU Designers `http://www.informit.com/articles/article.aspx?p=2142913`

[68] W. Gordon, "What Is HEVC H.265 Video, and Why Is It So Important for 4K Movies?", How-To Geek, 2018. [Online]. Available: https://www.howtogeek.com/342416/what-is-hevc-h.265-video-and-why-is-it-so-important-for-4k-movies/. [Accessed: 30- Nov- 2018].

[69] M. Budagavi, A. Fuldseth, G. Bjontegaard, V. Sze and M. Sadafale, "Core Transform Design in the High Efficiency Video Coding (HEVC) Standard", IEEE Journal of Selected Topics in Signal Processing, vol. 7, no. 6, pp. 1029-1041, 2013.

[70] B. N Madhukar, S. H Bharathi, "A New Property of the Discrete Cosine Transform-IV (DCT-IV)", Communication and Signal Processing (ICCSP) 2018 International Conference on, pp. 0144-0147, 2018.

[71] Mohammad Ali Pourabed, Sajjad Nouri, Jari Nurmi "Design and Implementation of 2-D IDCT/IDST-Specific Accelerator on Heterogeneous Multicore Architecture", 2018 IEEE Nordic Circuits and Systems Conference (NORCAS): NORCHIP and International Symposium of System-on-Chip (SoC)

[72] T. Ahonen and J. Nurmi, "Hierarchically Heterogeneous Network-on-Chip," EUROCON 2007 - The International Conference on "Computer as a Tool", Warsaw, 2007, pp. 2580-2586. doi: 10.1109/EURCON.2007.4400469

[73] C. Brunelli, F. Garzia, C. Giliberto and J. Nurmi, "A dedicated DMA logic addressing a time multiplexed memory to reduce the effects of the system bus bottleneck," 2008 International Conference on Field Programmable Logic and Applications, Heidelberg, 2008, pp. 487-490. doi: 10.1109/FPL.2008.4629990

[74] S. Nouri, W. Hussain and J. Nurmi, "Evaluation of a Heterogeneous Multicore Architecture by Design and Test of an OFDM Receiver," in IEEE Transactions on Parallel and Distributed Systems, vol. 28, no. 11, pp. 3171-3187, 1 Nov. 2017. doi: 10.1109/TPDS.2017.2706691

[75] R. Conceio, J. C. de Souza, R. Jeske, M. Porto, B. Zatt and L. Agostini, "Power efficient and high troughtput multi-size IDCT targeting UHD HEVC decoders," 2014 IEEE International Symposium on Circuits and Systems (ISCAS), Melbourne VIC, 2014, pp. 1925-1928. doi: 10.1109/ISCAS.2014.6865537

[76] L. Silva, F. Alves, J. Nacif, F. Passe, V. Vasconcelos and R. Ferreira, "CGRA HARP: Virtualization of a Reconfigurable Architecture on the Intel HARP Platform."

APPENDIX ALL CONTEXTS FOR IMPLEMENTATION

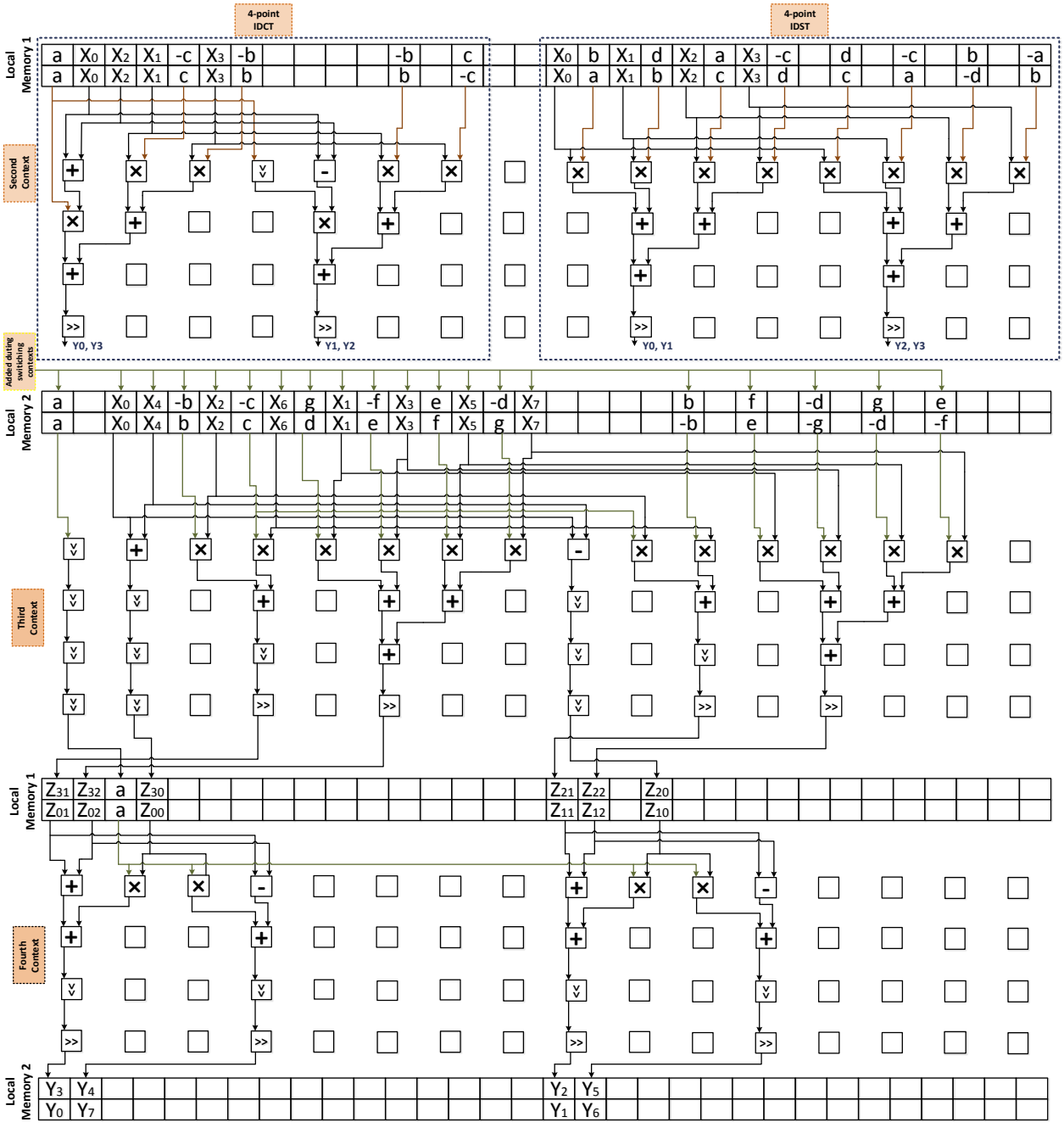

Figure 1 *Second, Third & Fourth Contexts for the Calculation of 4/8-point IDCT and 4-point IDST © 2018 IEEE [71]*